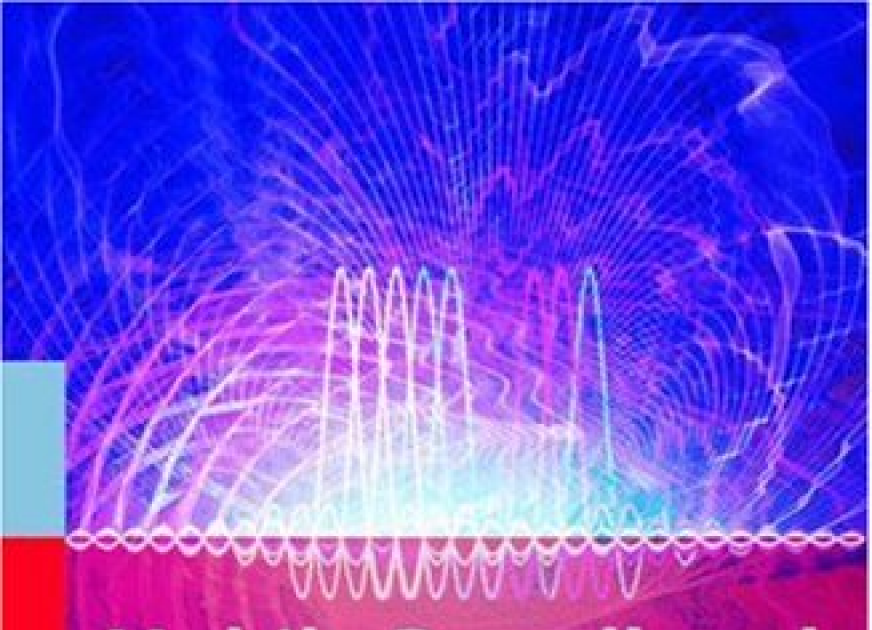


Mobile Broadband

Including WiMAX and LTE

Mustafa Ergen

Springer

Mobile Broadband

Including WiMAX and LTE

Mustafa Ergen

Mobile Broadband

Including WiMAX and LTE

Springer

Mustafa Ergen
Berkeley, CA
USA

ISBN: 978-0-387-68189-4 e-ISBN: 978-0-387-68192-4
DOI: 10.1007/978-0-387-68192-4

Library of Congress Control Number: 2008939013

While the advice and information in this book are believed to be true and accurate at the date of going to press, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

springer.com

Preface

This book attempts to provide an overview of IP-OFDMA technology, commencing with cellular and IP technology for the uninitiated, while endeavoring to pave the way toward OFDMA theory and emerging technologies, such as WiMAX, LTE, and beyond. The first half of the book ends with OFDM technology, and the second half of the book is targeted at more advanced readers, providing research and development-oriented outlook by introducing OFDMA and MIMO theory and end-to-end system architectures of IP- and OFDMA-based technologies.

The book comprises 13 chapters divided into three parts. Part I – constituted by Chaps. 1–3 – is a rudimentary introduction for those requiring a background in the field of cellular communication and All-IP Networking. Chapter 1 is introductory and is dedicated to discussing the history of cellular communications and the trend toward mobile broadband. Chapter 2 provides an overview of cellular communication with key insights to wireless challenges and features. Chapter 3 provides the same for IP networking.

Part II is comprised of Chaps. 4–7. Following an introduction to orthogonal frequency division multiplexing (OFDM) in Chap. 4, Chap. 5 is one of the core chapters of the book where orthogonal frequency division multiple access (OFDMA) is introduced in detail with resource allocation schemes. Chapter 6 talks about MIMO technologies and Chap. 7 introduces single-carrier frequency division multiple access (SC-FDMA) scheme – an OFDMA variant considered for uplink in LTE.

Part III, including Chaps. 8–13, introduces OFDMA-based access technologies. IEEE 802.16e-2005 based mobile WiMAX physical layer is described in Chap. 8, while IEEE 802.16e-2005 based mobile WiMAX medium access layer is detailed in Chap. 9. This is followed by Chap. 10, which concentrates on the networking layer specified by WiMAX Forum. Chapter 11 introduces air interface and networking framework of long-term evolution (LTE) out of Third Generation Partnership Project (3GPP), which is then followed by Chap. 12 that talks briefly about that of ultra mobile broadband (UMB) out of 3GPP2. In Chap. 13, we conclude the book with interworking solutions of access schemes presented earlier together with common IMS and PCC functions. In addition, we review future OFDMA-based technologies such as upcoming IEEE 802.16j and IEEE 802.16m for multihop relay and advanced air interface respectively as amendments to WiMAX. We then talk about IEEE 802.20 as a complement to UMB and cognitive radio-based IEEE 802.22 for wireless regional area networks.

The purpose of this book is to provide a comprehensive guide to researchers, engineers, students, or anyone else who is interested in the development and deployment of next generation OFDMA-based mobile broadband systems. The book targets to focus on a rapidly evolving area, and we have tried to keep it with up-to-date information. Despite the efforts to provide the text error free, for any errors that remain, comments and suggestions are welcome, which the will be used for preparing future editions. I can be reached via email at `ergen@cal.berkeley.edu`.

Finally, I thank my colleagues and my family for their constant support and patience. This book is dedicated to them.

Copyrighted material is reprinted with permission from IEEE Std 802.16. Permission is also granted for the use of IEEE Std 802.16j draft; IEEE Std 802.11n draft; and IEEE Std 802.16m working group documents. The IEEE disclaims any responsibility or liability resulting from the placement and use in the described manner.

Copyrighted material is reprinted with Permission of WiMAX Forum. “WiMAX,” “Mobile WiMAX,” “Fixed WiMAX,” “WiMAX Forum,” “WiMAX Certified,” “WiMAX Forum Certified,” the WiMAX Forum logo and the WiMAX Forum Certified logo are trademarks of the WiMAX Forum. The WiMAX Forum disclaims any responsibility or liability resulting from the placement and use in the described manner.

Copyrighted material is used under written permission of 3GPP TSs/TRs by ETSI. “LTE” is trademark of 3GPP. The 3GPP disclaims any responsibility or liability resulting from the placement and use in the described manner.

Copyrighted material is used under written permission of the Organizational Partners of the Third Generation Partnership Project 2 (3GPP2) and Telecommunications Industry Association. “UMB” is trademark of 3GPP2. The 3GPP2 disclaims any responsibility or liability resulting from the placement and use in the described manner.

Berkeley, CA *Mustafa Ergen*

Contents

Part I Fundamentals of Wireless Communication and IP Networking

Part II Theory of OFDMA and MIMO

Part III Applications of IP-OFDMA

Chapter 1
Introduction to Mobile Broadband

1.1 Introduction

A long way in a remarkably short time has been achieved in the history of wireless. Evolution of wireless access technologies is about to the reach its fourth generation (4G). Looking past, wireless access technologies has followed different evolutionary paths aimed at unified target: performance and efficiency in high mobile environment. The first generation (1G) has fulfilled the basic mobile voice, while the second generation (2G) has introduced capacity and coverage. This is followed by the third generation (3G), which has quest for data at higher speeds to open the gates for truly "mobile broadband" experience.[1]

What is "mobile broadband" then? Broadband refers to an Internet connection that allows support for data, voice, and video information at high speeds, typically given by land-based high-speed connectivity such as DSL or cable services. On the one hand, it is considered broad because multiple types of services can travel across the wide band, and mobile broadband, on the other hand, pushes these services to mobile devices.

We are seeing that mobile broadband technologies are reaching a commonality in the air interface and networking architecture; they are being converged to an IP-based network architecture with Orthogonal Frequency Division Multiple Access (OFDMA) based air interface technology. Although network evolution has not reached to the point of true and full *convergence*, wireless access networks, all at various stage of evolution, is being designed to support ubiquitous delivery of multimedia services via *internetworking*.

The transition to full convergence itself presents a set of unique challenges that the industry needs to address, however, IP-OFDMA-based technologies, the subject of this book, at one end and common policy control and multimedia services at the other end are good starts for full convergence.

[1] "Gartner predicts that mobile connections will top 3 billion worldwide by 2008 and that overall telecommunications services and equipment total revenue will reach $1.89 trillion (US) in 2009".

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_1, © Springer Science+Business Media LLC 2009

First worldwide debut of IP-OFDMA-based mobile broadband is with WiMAX (Worldwide Interoperability for Microwave Access) technology. This may be followed by Long Term Evolution (LTE), Ultra Mobile Broadband (UMB), and others. These standards are developed by partnership organizations and Internet Engineering Task Force (IETF[2], http://www.ietf.org). The Third Generation Partnership Project[3] (3GPP, http://www.3gpp.org) is responsible for LTE, while Third Generation Partnership Project 2[4] (3GPP2, http://www.3gpp2.org) deals with UMB. WiMAX is the exception to this since it is developed by WiMAX Forum (http://www.wimaxforum.org) and Institute of Electrical and Electronics Engineers (IEEE, http://www.ieee.org).

The underlying technology of WiMAX is considered to be a 4G system but early evolution and adoption of WiMAX has led the IEEE and the WiMAX Forum to ask R-ITU (Radiocommunication sector of the International Telecommunication Union) to include mobile WiMAX based on 802.16e into its IMT2000[5] specification (International Mobile Telecommunications 2000). WiMAX is included in IMT2000 in October 2007, which was originally created to harmonize 3G mobile systems. IMT2000 now supports seven different access technologies, including OFDMA (WiMAX), FDMA (Frequency Division Multiple Access), TDMA (Time Division Multiple Access), and CDMA (Code Division Multiple Access) as shown in Table 1.1. This will put OFDMA on a comparable worldwide footing with other recent and planned enhancements to 3G technology. As a result, alternative migration path as seen in Fig. 1.1 is now an option for operators to debut for value-added broadband services.

What remains for 4G then? IMT-Advanced, which is the ITU umbrella name for future 4G technologies has set vision of the characteristic of future 4G IMT-Advanced systems. Although there is no clear definition as of now, the ITU-R M.1645 considers a radio interface(s) that need to support data rates up to approximately 100 Mbps for high mobility such as mobile access and 1 Gbps for low

[2] "The Internet Engineering Task Force is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. An Internet document can be submitted to the IETF by anyone, but the IETF decides if the document becomes an RFC (Request for Comments), which has started in 1969 when the Internet was the ARPANET. Eventually, if it gains enough interest, it may evolve into an Internet standard. Each RFC is designated by an RFC number. Once published, an RFC never changes...".

[3] The 3GPP is formed by ETSI Europe, T1 USA, CWTS China, TTC Japan, ARIB Japan, TTA Korea.

[4] The 3GPP2 is formed by TIA USA, CWTS China, TTC Japan, ARIB Japan, TTA Korea.

[5] IMT2000 is particularly a framework that defines the criteria of ubiquitous support. The key criterias are:

- High transmission rates
- Fixed line voice quality
- Global roaming and circuit switched services support
- Multiple simultaneous services
- Increased capacity and spectral efficiency
- Symmetric and asymmetric transmission of data

Table 1.1 IMT2000

UMTS/WCDMA	CDMA Direct Spread
CDMA2000	CDMA Multi-Carrier
UMTS-TDD	Time-Code
TD-SCDMA	Time-Code
UWC-136	Single Carrier
IS-136	Single Carrier
EDGE	Single Carrier
DECT	FDMA/TDMA
WiMAX	OFDMA TDD

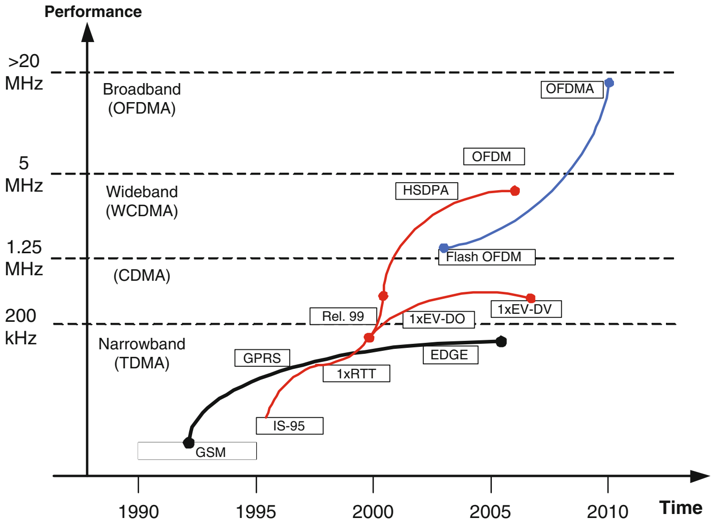


Fig. 1.1 Evolution of radio technologies source: Siemens

mobility such as nomadic/local access. These figures are seen to be the target and be researched and investigated further for feasible implementation. Current targeted landscape is shown in Fig. 1.2.

As can be seen mobile WiMAX based on 802.16e (We call WiMAX-e) would not qualify as a 4G IMT-Advanced standard since data rates even under ideal conditions are much lower but IEEE 802.16m, which is considered as the next Mobile WiMAX technology (we call WiMAX-m) and expected to be ratified in 2009, satisfies 4G requirements by achieving 1 Gbps data rate. Similar to current 802.16e Mobile WiMAX, the 802.16 m standard would use multiple-input, multiple-output (MIMO) antenna technology, while maintaining backward compatibility with the existing standards.

The speed on the order of 1 Gbps reportedly can be reached by using larger antenna arrays but current research shows that the data rate requirements described in ITU-R M.1645 can only be achieved with frequency bands above 100 MHz; however, there are very few large bands available. These requirements might be relaxed for the final release of 4G IMT-Advanced.

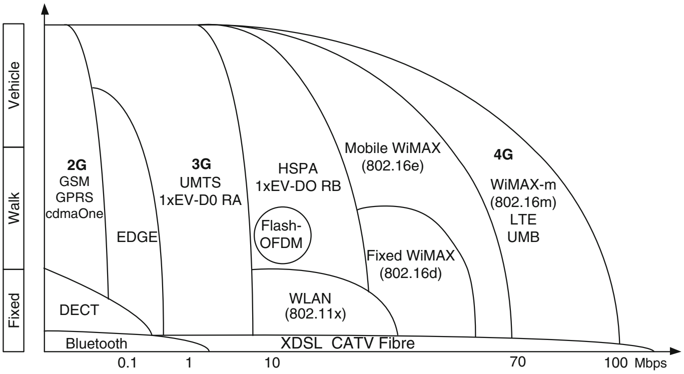


Fig. 1.2 Wireless standard landscape

We now start introducing the cellular evolution and broadband evolution in detail. First, we start with cellular systems that are introduced in the pre-3G era and also talk about the broadband services of that era. Later, we discuss the 3G cellular evaluation of 3GPP/2 and also introduce the broadband wireless access. At the last stage of the evolution, we talk about the motivation toward mobile WiMAX and 4G. Finally, we conclude the chapter with a discussion of key features and market of mobile broadband.

1.2 Before 3G and Broadband

Mobile broadband has two dimensions: mobility and broadband. However, traditionally, mobility first emerged for voice communication with cellular systems, and broadband has started with no mobility. Let us look first how these two have evolved to mobile broadband.

1.2.1 Cellular Communication

The most notable 1G cellular system was called the Advanced Mobile Phone System (AMPS), which was introduced by Bell Labs on the basis of cellular concept in 1947 and deployed worldwide in the 1980s. AMPS is an analog FDMA-based system for voice communication through 30 KHz FM modulated channels.[6] It is still being used in some rural areas of the United States however first generation cellular systems has lacked *uniform standardization*, which throttled the penetration.

[6] FCC has allocated 50 MHz total bandwidth for uplink and downlink.

Standardization has started with the 2G cellular systems. Global Systems for Mobile Communications (GSM) standard of Europe introduced digital communication with a combination of TDMA and slow frequency hopping for the voice communication. In the United States, 2G cellular standardization process at the 900 MHz followed two prong ways: Interim Standard-136 (IS) standard, evolved from IS-54,[7] considered TDMA and FDMA with phase-shift keyed modulation and cdmaOne IS-95 standard, first published in 1993, utilized direct-sequence CDMA with phase-shift keyed modulation and coding. In 2G, although standardization is present, a new challenge arose: *frequency allocation*.[8] The 2G standards are allowed in 2 GHz PCS (Personal Communications System) band but frequency band allocation in Europe is different from the one in the US, which made impossible to roam between systems nationwide or globally without a multimode phone.

The 2G has evolved to offer packet-based data services with GPRS (General Packet Radio Service) and EDGE (Enhanced Data rates for GSM Evolution) within GSM systems. GPRS reached peak data rates up to 140 Kbps when a user aggregates all timeslots. EDGE has increased data rates up to 384 Kbps with high-level modulation and coding. Adaptive Modulation and Coding (AMC) is introduced by EDGE to adaptively select the best modulation according to the received Signal-to-Noise-Ratio (SNR) feedback. IS-95A provided circuit-switched data connections at 14.4 Kbps and IS-95B[9] systems has offered 64 Kbps packet-switched data, in addition to voice services.

1.2.2 Broadband and WLAN/WiFi

Another evaluation is as we said broadband connectivity, which has started with Digital Subscriber Line (DSL) and cable modem technology. DSL utilizes the twisted pair copper wire of the local loop of the public switched telephone network (PSTN), which is used to carry Plain Old Telephone Service (POTS) voice communication between 300 and 3.4 KHz. DSL uses the bandwidth beyond 3.4 KHz. The length and quality of the loop determines the upper limit that can be utilized for DSL connection. DSL utilizes Discrete Multitone Modulation (aka Orthogonal Frequency Division Multiplexing (OFDM)) and DSL modem converts digital data

[7] It is the first digital 1G cellular system over TDMA. Also, called Digital-AMPS.

[8] Spectrum allocation and controlling use is governed by government agencies. Federal Communications Commission (FCC) regulates the commercial use and Office of Spectral Management (OSM) regulates the military use in the United States. European Telecommunications Standard Institute (ETSI) regulates the spectrum in Europe and International Telecommunications Union (ITU) governs globally. Frequency bands could be licensed or license-exempt. Band for licensed use is determined through spectrum auctions and primary purpose of license-exempt operation is to encourage innovation and low-cost deployment.

[9] The IS-95B revision, also termed TIA/EIA-95, combines IS-95A, ANSI-J-STD-008, and TSB-74 standards into a single document. The ANSI-J-STD-008 specification, published in 1995, defines a compatibility standard for 1.8–2.0 GHz CDMA PCS systems. TSB-74 describes interaction between IS-95A and CDMA PCS systems that conform to ANSI-J-STD-008.

into analog waveform. These waveforms coming from various DSL modems are aggregated at a Digital Subscriber Line Access Multiplexer (DSLAM), which acts as a gateway to other networking transports. *DSL Forum* has driven global standardization with several xDSL standards such as ADSL, SHDSL, VDSL, ADSL2plus, VDSL2, and more. ADSL is holding more than 60% of the broadband subscribers, which was around 350 million worldwide at the end of 2007. ADSL standard can deliver 8 Mbps to the customer over about 2 km. The latest ADSL2plus can go up to 24 Mbps depending on the distance from the DSLAM since increasing the distance to DSLAM decreases the performance. The first DSL debut was for Internet connection, lately it has been converging to provide bundled services like voice, video especially Internet Protocol Television (IPTV), and data.

The cable modem technology comprises several standards to deliver high-speed data transfer over an existing coaxial Cable TV (CATV) system. The *CableLabs* founded in 1988 by cable operation companies defines DOCSIS (Data Over Cable Service Interface Specification), which is an interface requirements for cable modems that are used in data transmission. Another standard from CableLabs is PacketCable built over DOCSIS to define interface specifications for delivering advanced, real-time multimedia services via IP technology. This includes multimedia services, such as IP telephony, multimedia conferencing, interactive gaming, and general multimedia applications. CableLabs also introduces Video on Demand (VoD) Metadata project to define specifications how the content package may be delivered from multiple content providers sent over diverse networks to cable operators. Lately, the CableHome project is introduced to extend high-quality cable-based services to network devices within the home to deliver voice, video especially high-definition TV (HDTV), and data.

The broadband is also evolving with xDSL and cable variants as well as new technologies like FTTH (fiber-to-the-home) over an optical fiber, which run directly onto the customer's premises unlike fiber-to-the-node (FTTN), fiber-to-the-curb (FTTC), or hybrid fibre-coaxial (HFC), all of which depend upon more traditional methods such as copper wire or coaxial cable for "last mile" delivery.

However, the broadband over DSL and cable are only capable to provide last mile connection with no mobility. Limited mobility is introduced with the introduction of Wireless Local Area Networking (WLAN) within the past decade. WLAN systems are confined to deliver wireless connectivity within a small range, and they are utilized to distribute fixed broadband connectivity to nomadic wireless users as well as users with pedestrian speed.

WLAN establishes wireless connection between wireless stations (such as PCs, laptops, handhelds, etc.) and the access point that connects to DSL or Cable modem or Ethernet for broadband connectivity. WLAN operates in unlicensed frequency bands. The primary unlicensed bands are the ISM (Industrial, Scientific, and Medical) bands at 900 MHz, 2.4 GHz, and 5.8 GHz and the Unlicensed National Information Infrastructure (U-NII) band at 5 GHz. WLAN is hosted in ISM band as secondary user and has to vacate if primary users are active. However, U-NII band does not have primary users.

The WLAN has been standardized in IEEE within 802.11 framework. The first standard 802.11b is introduced in 2.4 GHz ISM band for 83.5 MHz spectrum. The 802.11b utilized direct-spread spectrum to offer data rates up to 11 Mbps within 100m range. Later, IEEE 802.11a is introduced in 300 MHz of 5 GHz U-NII band. The 802.11a is the first standard in the wireless domain to use OFDM modulation to provide up to 54 Mbps within less than 100 m range. IEEE 802.11a has also more channels than 802.11b and has the ability to accommodate users with higher data rates. To leverage this system design, later IEEE 802.11g is introduced in the 2.4 GHz band with the same design as in IEEE 802.11a. IEEE 802.11g is designed also to be backward compatible with IEEE 802.11b. These systems, although evolved to support higher rates, lack a MAC protocol that supports Quality of Service (QoS). Later, IEEE 802.11e framework addressed QoS and IEEE 802.11n framework is designed to accommodate MIMO technology with OFDM modulation. In Europe, HiPERLAN (High Performance Radio LAN) standards are designed to introduce WLAN service. The HiPERLAN/2 standard also utilizes OFDM standard as in IEEE 802.11a in 5 GHz U-NII band.

WLAN standard within IEEE frame only defines the physical and MAC layers. The industry formed the *Wi-Fi Alliance* as a nonprofit industry association to enhance the user experience by defining the networking layer as well as testing and certification programs. Currently, wireless LAN is proliferating at homes, enterprises, and even in cities, and has become the standard for “last feet” broadband connectivity. The success of WLAN has accelerated the hype toward broadband wireless access with more mobility and guaranteed QoS.

1.3 3G and Broadband Wireless

Moving toward mobility and high speed from broadband and cellular systems has continued in different angles in the third generation era. The 3GPP and 3GPP2 have introduced the 3G technologies as an evolution to their existing second generation paths. After summarizing these technologies, we give the evolution of broadband to WiMAX from broadband wireless access.

1.3.1 The 3GPP Family

Universal Mobile Telecommunications System (UMTS), which is based on Wideband Code Division Multiple Access (WCDMA), has been studied in Release-1999 (Rel-99) of 3GPP and published in 2000. UMTS was the next step after GSM, GPRS, and EDGE to offer improved voice and data services with a 5 MHz bandwidth. Rapid growth of UMTS, where future projection is seen in Table 1.2, has led to the next step in evolutionary phase termed, Release-2005 (Rel-5).

Table 1.2 Global UMTS customer forecast by World Cellular Information Service, Informa Telecoms and Media, May 2007

2007	200M
2008	350M
2009	500M
2010	700M
2011	900M
2012	1250M

Rel-5 provided High Speed Downlink Packet Access (HSDPA) that brought spectral efficiency for higher-speed data services. Rel-5 also introduced IP Multimedia Subsystem (IMS) and IP UMTS Terrestrial Radio Access Network (UTRAN) to offer flexibility to operator to provide such hosted services for greater user experience. Meanwhile, Rel-4 is introduced in March 2001, which separated call and bearer in the core network.

On the one hand, Rel-6, introduced in March 2005, came with High Speed Uplink Packet Access (HSUPA), Multimedia Broadcast Multicast Service (MBMS), and advanced receivers. The combination of HSDPA and HSUPA is called HSPA.

Rel-7, on the other hand, focuses on MIMO technology and flat-IP based base stations. GPRS Tunneling Protocol (GTP) has started to be used in order to connect packet switched network to radio access network. Rel-7 is expected to finish in 2008 with new enhancements and it is termed HSPA Evolution, commonly known as HSPA+. Rel-7 has also improved receiver architecture and brought interference aware receivers (referred as type 2i and type 3i, which are extensions to existing type 2 and type 3 receivers). The receiver employs interference aware structure, which not only takes into account the channel response matrix of the serving cell but also the channel response matrix of the interfering cell that has the most significant power. Rel-7 also introduced the use of higher order modulations such as 64QAM with MIMO support since in Rel-6, HSPA systems used 16QAM in the downlink and QPSK in the uplink. To reduce latency when exiting the idle mode, Continuous Packet Connectivity (CPC) has been introduced for data users. This mainly keeps more users in the cell active state. The protocol is modified to ensure the user keep synchronized and the power control ready for rapid resumption (Table 1.3).

In the network side, architecture has been improved as well. HSPA+ has integrated the RNC (Radio Network Controller) to NodeB (base station) to reduce latency and to make the architecture flatter and simpler. It is also a good move toward femtocell[10] deployments and a good step to enable packet-based services toward LTE since HSPA+ is considered to be the "missing link" between HSPA and LTE.[11]

[10] "Femtocells are being standardized in the *Femto Forum* (`http://www.femtoforum.org`) as a low-power wireless access points that operate in licensed spectrum to connect standard mobile devices to a mobile operator's network using residential DSL or cable broadband connections...".

[11] Rel-7 also introduced enhancements in device perspective. Single public identity has been provided to IMS user with multiple device support. Mobile payment or transportation applications has been addressed with Universal Integrated Circuit Card (UICC), collaborated with OMA (Open

Table 1.3 Data speed of various technologies:

Technology	Bandwidth	Technology	DL/UL peak
WCDMA Rel. 99	5 MHz FDD	TDM/CDMA	384/384 Kbps
HSPA Rel. 6	5 MHz FDD	TDM/CDMA	1.8–14.4/5.72 Mbps
HSPA+ Rel. 7	5 MHz FDD	TDM/CDMA	22/11 Mbps
LTE	1.25–20 MHz FDD	OFDMA/SC-FDMA	100/50 Mbps
CDMA2000 1x	1.25 MHz FDD	TDM/CDMA	153/153 Kbps
1xEV-DO Rev-0	1.25 MHz FDD	TDM/CDMA	2.4 Mbps/153 Kbps
1xEV-DO Rev-A	1.25 MHz FDD	TDM/CDMA	3.1/1.8 Mbps
1xEV-DO Rev-B	5 MHz FDD	TDM/CDMA	14.7/5.4 Mbps
UMB	1.25–20 MHz FDD	OFDMA	33-152/17-75 Mbps
WiFi	20 MHz TDD for 802.11a/g	CSMA/OFDM	54 Mbps shared
Fixed WiMAX	TDD, FDD 3.5 MHz, 7 MHz, 10 MHz	TDM/OFDM	9.4/3.3 Mbps with 3:1; 6.1/6.5 Mbps with 1:1
Mobile WiMAX	TDD 3.5 MHz, 7 MHz, 5 MHz, 10 MHz, 8.75 MHz	TDM/OFDMA	46/7 Mbps 2×2 MIMO in 10 Hz with 3:1; 32/4 Mbps with 1:1

HSPA operates in 800, 900, 1,800, 1,900, 2,100 MHz; EV-DO operates in 800, 900, 1800, 1,900 MHz; WiFi operates in 2.4 GHz, 5 GHz; fixed WiMAX operates in 3.5 GHz, and 5.8 GHz (unlicensed) initially; mobile WiMAX operates in 2.3 GHz, 2.5 GHz, and 3.5 GHz initially. The 3:1 and 1:1 stands for DL:UL ratio in TDD mode

1.3.2 The 3GPP2 Family

The 3GPP2 has continued to evolve its second generation (IS-95) based systems with EV-DO (Evolution-Data Optimized) series of CDMA2000 standard.[12] First standard of series, termed CDMA2000 1xEV-DO, introduces data-centric broadband network to deliver data rates beyond 2 Mbps in a mobile environment. In 2001, CDMA2000 1xEV-DO was approved as an IMT2000 standard as CDMA2000 High Rate Packet Data (HRPD) Air Interface, IS-856. CDMA2000 1xEV-DO Release 0 (Rel-0) offers high-speed data access up to 2.4 Mbps and it was the first mobile broadband technology deployed worldwide.[13]

Rel-0 provides a peak data rate of 2.4 Mbps in the forward link (FL) and 153 Kbps in the reverse link (RL) in a single 1.25 MHz FDD (Frequency Division Duplexing) carrier. In commercial networks, Rel 0 delivers average throughput of 300–700 Kbps in the forward link and 70–90 Kbps in the reverse link. Rel-0 has also

Mobile Alliance) and ETSI-SCP. Smart Card Server located in UICC offers secure and portable contactless exchanges with the Single Wire Protocol.

[12] "The CDMA2000 standards CDMA2000 1xRTT, CDMA2000 EV-DO, and CDMA2000 EV-DV are approved radio interfaces for the ITU's IMT-2000 standard. CDMA2000 is a registered trademark of the Telecommunications Industry Association (TIA-USA) in the United States, not a generic term like CDMA. CDMA2000 is defined to operate at 450, 700, 800, 900, 1,700, 1,800, 1,900, and 2,100 MHz. Source: Wikipedia".

[13] South Korea adopted first in 2002.

started "always on" user experience as in IP and also supports IP-based network connectivity and applications. CDMA2000 1xEV-DO devices include a CDMA2000 1X modem in order to be compatible with CDMA2000 1X and cdmaOne systems.

In addition to the air interface techniques of CDMA2000 1X, the following new high-speed packet data transmission enhancements are incorporated into Rel-0: downlink channelization to offer higher rate with bundling, Adaptive Modulation and Coding, Hybrid-ARQ, etc.

CDMA2000 1xEV-DO Revision A (Rev-A) is an evolution of CDMA2000 1xEV-DO Rel-0 to increase peak rates on reverse and forward links to support a wide-variety of symmetric, delay-sensitive, real-time, and concurrent voice and broadband data applications. It also incorporates OFDM technology to enable multicasting (one-to-many) for multimedia content delivery. Rev-A has introduced first All-IP based broadband architecture in 2006 to support time-sensitive applications such as VoIP, etc. Rev-A provides a peak data rate of 3.1 Mbps in the forward link and 1.8 Mbps in the reverse link with a 1.25 MHz FDD carrier. However, in commercial networks, Rev–A achieves average throughput of 450–800 Kbps in the forward link and 300–400 Kbps in the reverse link.

As the successor of Rev-A, CDMA2000 1xEV-DO Revision B (Rev-B) introduces dynamic bandwidth allocation to provide higher performance by aggregating multiple 1.25 MHz Rev-A channels. Consequently, peak data rates scales with the number of carriers aggregated. When 15 channels are combined within a 20 MHz bandwidth, Rev-B delivers up to 46.6 Mbps in the forward link and 27 Mbps in the reverse link. However, with 5 MHz aggregation, the peak data rates are around 14.7 Mbps.[14] Rev-B also supports OFDM based multicasting and introduces lower latency for delay sensitive applications.

1.3.3 Broadband Wireless Access

Broadband Wireless Access (BWA) has started with a fixed access in mind to compete with DSL and cable modem since rapid growth of broadband access has created demand for new wireless technologies to reduce the cost of operation and by pass monopoly of service providers in wire-line access. We give a chronological listing of BWA toward fixed WiMAX in this section and mobile WiMAX in the next section.

The Local Multipoint Distribution Systems (LMDS) is the first notable BWA that showed a short-lived rapid success as a wireless alternative to fiber and coaxial cables in the late 1990s. LMDS has utilized 28 & 31 GHz with two types of

[14] "With the 64QAM scheme, the peak data rate in the forward link increases in a single 1.25 MHz carrier to 4.9 Mbps however an aggregated 5 MHz will deliver up to 14.7 Mbps and within 20 MHz of bandwidth, it is up to 73.5 Mbps...".

LMDS licenses to or in? Offer up to several hundreds of megabits per second. However, LMDS system requires roof-top antennas to achieve line-of-sight (LOS) connection.

Multichannel Multipoint Distribution Services (MMDS or Wireless Cable) technology has emerged at 2.5 GHz and become popular in sparsely populated rural areas. LMDS and MMDS have adapted the modified version of DOCSIS for wireless broadband also known as DOCSIS+. MMDS provided greater range than LMDS but still required LOS link to operate.

The LOS challenge of broadband wireless has tackled with OFDM modulation and standardization activities have begun in 1998 by IEEE under the 802.16 working group. This group has targeted to standardize the technology for Wireless Metropolitan Area Network (Wireless MAN), also adopted by ETSI HiPERMAN (High Performance Radio Metropolitan Area Network). In 2001, first standard is approved as Wireless MAN-SC that specifies a single-carrier technology for operation in the 10–66 GHz band like LMDS. Non-LOS (NLOS) has been addressed in 2–11 GHz band for licensed and unlicensed frequencies as amendments to existing 802.16 standard. The IEEE 802.16a, completed in 2003 introduced three access schemes: single-carrier, OFDM and OFDMA for fixed NLOS access. It also specifies a common MAC layer for all three access schemes where concepts were mainly adapted for wireless from DOCSIS. The IEEE 802.16-2004 standard ratified in 2004 replaced IEEE 802.16, 802.16a, and 802.16c standards with a single standard and formed the basis for fixed WiMAX solution. In 2005, IEEE 802.16e-2005 amendment, which forms the basis for mobile WiMAX, is ratified to introduce enhancements for high-speed mobility. The IEEE 802.16 framework specifies the physical and media access control layers but does not deal with the end-to-end systems' requirements and interoperability criteria of systems built on these requirements. The industry-led *WiMAX Forum* was organized to fill this void to address fixed WiMAX and mobile WiMAX network architectures and protocols including interoperability and certification.

Currently, WiMAX Forum introduced two system profiles: fixed system profile based on IEEE 802.16-2004 OFDM physical layer, and mobile system profile based on IEEE 802.16e-2005 scalable OFDMA physical layer. Besides system profile, certification profiles are defined to specify the operating frequency, channel bandwidth, and duplexing mode as seen in Tables 1.4 and 1.5.

Table 1.4 Fixed WiMAX initial certification profiles

Band (GHZ)	Channel Bandwidth (MHz)	OFDM FFT size	Duplexing
3.5	3.5	256	FDD
3.5	3.5	256	TDD
3.5	7	256	FDD
3.5	7	256	TDD
3.5	10	256	TDD

Table 1.5 Release-1 System Profiles for Mobile WiMAX

Channel BW (MHz)	FFT size	2.3–2.4 GHz	2.305–2.32, 2.345–2.36 GHz	2.496–2.69 GHz	3.3–3.4 GHz	3.4–3.8 GHz
1.25	128					
5.0	512	TDD	TDD	TDD	TDD	TDD
7.0	1024				TDD	TDD
8.75	1024	TDD				
10	1024	TDD	TDD	TDD	TDD	TDD
20	2048					

FDD mode is being designed. WiBro, Mobile WiMAX in Korea, operates in 2.3 GHz band with 9 MHz channel spacing in IEEE 802.16e-2005 TDD mode

1.4 Mobile WiMAX and 4G

Mobile WiMAX has evolved from fixed wireless access and inherits its features for optimized broadband data services. EV-DO and HSPA, 3G CDMA standards, have been originally conceived for mobile voice services and inherit both advantages and limitations of legacy 3G systems. Consequently, mobile WiMAX faces the challenge to support mobility whereas 3G systems faces the challenge to support higher data rates.

Mobile WiMAX provides higher data rates with OFDMA support and introduces several key features necessary for delivering mobility at vehicular speeds with QoS comparable to broadband access alternatives. Several features that are used to enhance data throughput are common to EV-DO and HSPA: Adaptive Modulation and Coding (AMC), Hybrid-ARQ, fast scheduling, and bandwidth efficient handover. The key differenc is in duplexing where EV-DO and HSPA are FDD operating on a carrier frequency of 2.0 GHz, whereas mobile WiMAX is currently TDD (Time Division Duplexing) operating at 2.5 GHz. Mobile WiMAX has higher tolerance to multipath and self-interference and provides orthogonal uplink multiple access with frequency selective scheduling and fractional frequency reuse.

Unlike EV-DO and HSPA, Mobile WiMAX is also capable of utilizing 2×2 MIMO in addition to 1×2 SIMO. Performance comparison has shown that in 10 MHz channel, mobile WiMAX has a net downlink throughput by 9–14 Mbps with MIMO and 6–9 Mbps with SIMO per channel/sector as compared to $\sim$4 Mbps with EV-DO Rev-B and HSPA. This leads to a downlink spectral efficiency around 1.9 bps/Hz with MIMO at maximum when compared with 0.8 bps/Hz with EV-DO Rev-B and HSPA. Consequently, fewer base stations are required to achieve the desired data density.

There are already other contenders for mobile broadband besides WiMAX as seen in Fig. 1.3: Long Term Evolution (LTE) (Release 8) out of 3GPP and Ultra Mobile Broadband (UMB) (formerly CDMA2000 1xEV-DO Rev-C) out of 3GPP2. The good news is that they are being designed with the same air interface (OFDMA) as WiMAX. Change from WCDMA to OFDMA will be the second significant change

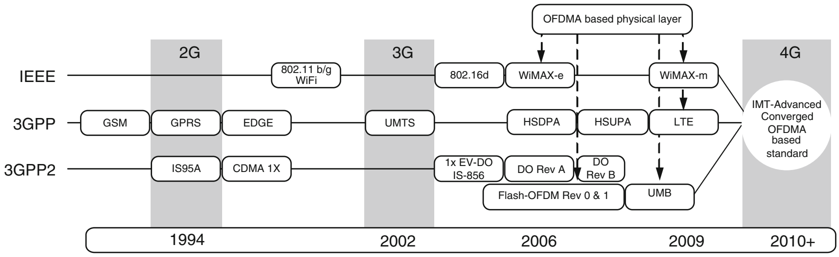


Fig. 1.3 Evolutionary path of cellular technology

in 3GPP standards after TDMA to WCDMA change during the shift from 2G to 3G systems. OFDMA selection is driven by the demand for higher spectral efficiency and low cost per bit since the basic problem for a service provider is to get more data to users, quicker and cheaper. Because WCDMA has a restriction to scale in bandwidth, OFDMA is selected. OFDMA solves this problem by splitting the high-speed data stream into several lower speed data streams and sending the lower speed streams on individual frequency channels. In the receiver, the user recombines these lower streams to construct a high-speed data stream.

Besides OFDMA technology, all three are based on IP services with no backward compatibility for circuit-switched services. This is another real break in technologies when moving to 4G since it gives a significant advantage to technologies that are coming out of blue like WiMAX. Operators now have a choice thereby they need not to follow the evolution path of 2G or 3G standard that they are currently using. WiMAX is also seen as the only player that can offer a unified fixed-mobile solution in broadband wireless as well as mobile broadband markets.

In brief, there are certainly similarities and few differences in the technology: performance, time-line, cost of operation, and IPR[15] are ingredients to determine a selection for mobile operators with regard to what the ecosystem is like and what the mobile community as a group wants to do.

1.5 Key Features

From technical perspective, fundamental goal of mobile broadband is to offer higher data rates with reduced latency. The key characteristics of a typical mobile broadband system are summarized here:

- *Increased data rates:* OFDMA based air interface is the key technology to offer higher data rates with higher order modulation schemes such as 64QAM, and

[15] "The patents and other intellectual property is one of the key requirements of technological and market development. WiMAX and other wireless technologies are built on the accomplishments of thousands. A favorable patent regimen with lower cost and converging Intellectual Property Rights (IPR) may foster the technology...".

sophisticated FEC (Forward Error Correction) schemes such as convolutional coding, turbo coding, alongside complementary radio techniques like MIMO and beamforming with up to four antennas per station.

- *High spectral efficiency:* Operators seek to increase the number of customers within their existing spectrum allocations, with reduced cost of per bit.
- *Flexible radio planning:* Deployment flexibility gives operators to change the cell size depending on the demand.
- *Reduced latency:* Next generation applications requires reduced round-trip times to 10 ms or even less. Responsiveness enables interactive, real-time services such as high-quality audio/videoconferencing and multi player gaming.
- *All-IP architecture:* Transition to a "flat", all-IP based core network will enable PC-like services such as voice, video, data and improves the interworking to other fixed and mobile networks.
- *Interworking:* Mobile broadband requires interworking to existing technologies to support fixed-mobile convergence.
- *Open interfaces:* Open interfaces enable multi-vendor network operation to give operator great flexibility to select best solutions. This leverages developments in other industries including Internet, PC, and network systems, etc.
- *Spectral flexibility:* Scalable bandwidths give operators flexibility to reuse their existing spectrum allocations. This is called "refarming" in the mobile telecommunications value chain as a cost-efficient option to address increasing traffic demands.
- *Cost reduction capabilities:* New features like Mobile Virtual Network Operation (MVNO), network sharing, or self optimizing networks are needed to reduce the OPEX (OPerational EXpenditure).
- *Support for data centric services:* Operators are looking for solutions to revert their declining ARPU (Average Revenue Per User).

1.6 Mobile Broadband Market

In the near future, OFDMA-based mobile broadband with the recent progress made by technical specifications and vendor technology demonstrations will emerge as successor to cellular systems as a broadband wireless solution.

Higher data rates and higher spectral efficiency become imminent with growing demands for wireless data services. 3G networks, which are being deployed worldwide, demonstrated a good example for an operator to increase their ARPU from broadband data services. Cost of network together with spectral cost will determine how far the current existing 3G networks advocate the current rise in ARPU with data services. In parallel, new services are created to boost the consumer demand such as mobile content, entertainment, advertising, MMS, video, etc. The most important part that will drive the convergence is content created by user (UGC), which will make "ease of use" as the next "big thing" in terms of technology.

Current GSM mobile subscriber rate is a good indication for the potential market for wireless broadband data. Besides ADSL and cable connections for broadband, GSM/UMTS mobile subscribers worldwide is expected to increase to four billion in 2011. Operator is already seeing increase in data ARPU and decrease in voice ARPU. Moreover, according to a study by the Online Publishers Association, currently more than 76% of mobile phones are Web-enabled in US and Europe in addition to PC cards and embedded modems.

The success of mobile broadband will also be driven by the development of user-friendly handsets and applications including mobile music, multimedia messaging, gambling, and mobile TV. The IP Multimedia Subsystem (IMS) will play a key role in adoption of mobile broadband with its ability to offer applications of wireline word via wireless by supporting more than one access networks, including WiMAX, LTE, UMB, CDMA2000, WLAN/WiFi, cable, xDSL, etc.

1.7 Summary

This chapter charts the technological roadmap for mobile broadband, backed by IP-OFDMA based convergence. Clear contenders for mobile broadband follows two evolutionary paths: from broadband access or from cellular communication. Key highlights of the chapter are as follows:

- OFDMA is selected as the air interface of various standards by the standardization bodies.
- Cellular networking has been moving toward a flat IP-based architecture for the past decade to offer simpler and scalable design.
- Inclusion of Mobile WiMAX in IMT2000 offers a 3G alternative to operators to migrate to an OFDMA-based system.
- Scalable OFDMA-based interface can be deployed in a variety of spectrum bands including 2.3, 2.5, 3.5 GHz as well as existing 3G spectrum.
- WiMAX MAC inherits features from DOCSIS, cable standard.
- There are clear contenders to Mobile WiMAX: LTE out of 3GPP and UMB out of 3GPP2.
- WiMAX-m based on IEEE 802.16 m is being designed to debut for 4G along with LTE and UMB with regard to the IMT-Advanced 4G criteria.
- Mobile WiMAX is the only player to address both cellular and fixed broadband.

References

1. Bolton, W., Xiao, Y., Guizani, M., "High802.20: mobile broadband wireless access," *IEEE Wireless Communications*, vol. 14, pp. 84–95, 2000.
2. Buckley, S., "WiMAX Forum gears up for mobility and more," *Horizon House Publications*, 2005.

3. Seybold, A.M., “WiMAX Again?” *Outlook Mobility Newsletter,* 2004.
4. Jackson, D., “Motorola announces plans to converge WiMAX and 4G,” *Primedia, Inc.*, 2005.
5. Walko, J., “Samsung Demos Korean Version of WiMAX At 4G Forum,” *Personal Tech pipeline eNewsletter, CMP Media LLC,* 2005.
6. Seals, T., “ WiMAXimum Exposure-A new type of Broadband wireless gathers momentum,” *Infrastructure Solutions,* 2004. `http://www.xchangemag. com.`
7. Paolini, M., “WiFi, WiMAX and 802.20-The Disruptive potential of wireless Broadband,” Senza Fili Consulting & BWES Ltd, 2004.
8. Haslam, D., “Providers reveal broadband wireless access deployment barriers in Sage research study,” 2002. `http://www.sageresearch.com`
9. Paolini, M., “WiFi When, where, and WiMAX,” 2004. `http://www.edn.com/article/CA419563.html.`
10. Fitchard, K., “WiMAX prepare to come of age,” 2005. `http://www.telephony-online.com/mag/.`
11. Thinkquest, “Wireless Communication Technologies- WiMAX,” `http://library.thinkquest.org/040i/01721/wireless/wimax.htm.`
12. “CDMA2000 1xEV-DO Revision A: The Gateway to True Mobile Broadband Multimedia,” CDMA Development Group, August 2006. `http://www.cdg.org/.`
13. Callahan, P., “Mobile VoIP Over 1xEV-DO,” Airvana, July, 2006.
14. Andrews, J. G., Ghosh, A., Muhammed, R., *Fundamentals of WiMAX,* Prentice Hall, 2007.
15. Goldsmith, A., *Wireless Communications,* Cambridge University Press, Cambridge, 2005.
16. Correia, L. M., *Mobile Broadband Multimedia Networks: Techniques, Models and Tools for 4G,* Academic Press, 2006.
17. CDMA2000 Technologies, CDMA Development Group. `http://www.cdg.org/.`
18. WiMAX Forum, “Executive Summary: Mobile WiMAX Performance and Comparative Summary,” Sept. 2006. `http://www.wimaxforum.org/.`
19. WiMAX Forum, “Mobile WiMAX Part I: A Technical Overview and Performance Evaluation,” 2006. `http://www.wimaxforum.org/.`
20. WiMAX Forum, “Mobile WiMAX Part II: A Comparative Analysis,” 2006. `http://www.wimaxforum.org/.`
21. WiMAX Forum, “KT Corporation to Launch Commercial WiBro Services in Mid-2006 Press Release,” Nov. 14, 2005. `http://www.wimaxforum.org/.`
22. 3GPP TSG-RAN-1, “Effective SIR Computation for OFDM System-Level Simulations,” R1-03-1370, Meeting #35, Lisbon, Portugal, November 2003. `http://www.3gpp.org/.`
23. 3GPP TSG-RAN-1, “System-Level evaluation of OFDM - further Considerations,” R1-031303, November 17-21, 2003. `http://www.3gpp.org/.`
24. 3GPP2 C.R1002-0, “CDMA2000 Evaluation Methodology,” December 2004. `http://www.3gpp2.org/.`

Chapter 2
Basics of Cellular Communication

Quotation from FCC's Statement is as follows;

"We are still living under a spectrum "management" regime that is 90 years old. It needs a hard look, and in my opinion, a new direction. Historically, I believe there have been four core assumptions underlying spectrum policy: Unregulated radio interference will lead to chaos;

- Spectrum is scarce.
- Government command and control of the scarce spectrum resource is the only way chaos can be avoided.
- The public interest centers on government choosing the highest and best use of the spectrum.

Today's environment has strained these assumptions to the breaking point. Modern technology has fundamentally changed the nature and extent of spectrum use. So the real question is, how do we fundamentally alter our spectrum policy to adapt to this reality? The good news is that while the proliferation of technology strains the old paradigm, it is also technology that will ultimately free spectrum from its former shackles..."

2.1 Cellular Concept

The ultimate objective of wireless communication is to host large number of users in a wide coverage. But as quoted above from Federal Communications Commission's statement that the *spectrum is scarce*. This limits coverage on the expense of number of users or vice versa.

Initial deployment of wireless networks has dated back to 1924 with one base station providing a city-wide coverage. Although achieving very good coverage, the network can only host a few users simultaneously. Another base station using the same spectrum and serving the same area cannot be placed since that would result in interference.

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_2, © Springer Science+Business Media LLC 2009

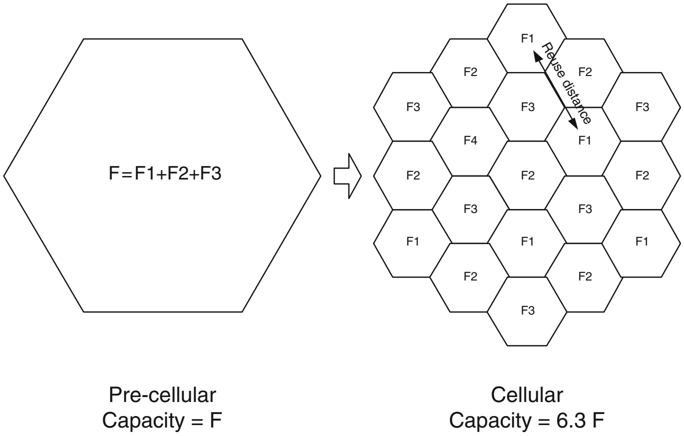


Fig. 2.1 Cellular concept

The cellular concept has introduced smaller cells operating with a channel, which is a split of the allocated spectrum. Number of base station is increased to achieve larger coverage and in order to reduce interference, using the same channel is not allowed in adjacent base stations but same channel is reused in other base stations that are spatially separated. Hence, the degree of spatial separation directly affects capacity and interference as seen in Fig. 2.1.

A cell can host limited number of users and to increase the capacity, if there is more demand, more number of base stations can be deployed with reduced coverage. Channels can be allocated with distributed fashion with spatial separation in mind for the same channels. For instance, if the allocated spectrum is F. F can be split into n channels. n channels is distributed to N base stations (BS). This is called cluster and cluster is replicated m times to cover the area. Total capacity C is then equals to $m \times F$. For instance, in precellular concept, total capacity is F since $m = 1$ and $n = 1$.

Of course, the above analysis gives theoretical capacity since in real deployment, cells operating with the same channel cause *co-channel interference* to each other. To reduce the co-channel interference, cells operating in the same channel should be separated by a distance to provide ample protection. Co-channel reuse ratio is given by D/R, where D is the distance of two same channel cells and R is cell radius.

There is also *adjacent channel interference*, which is basically a leak from adjacent channel in the spectrum due to imperfection in the devices. Adjacent channel interference can be minimized by keeping the frequency separation between each channel in a given cell as large as possible. Interference is further mitigated by controlling the power of mobile subscriber. Power control maintains the mobile transmission power low enough to maintain a good quality link. Mobile subscriber close to BS is forced to reduce the power and away from BS is forced to increase the transmit power.

2.1.1 Handover

Of course, in mobile networking with cellular deployment crossing multiple cells on the move is inevitable. Hence, the serving base station (BS) changes with mobility. Also, note that serving BS might change depending on the load conditions as well in which MS is involuntarily shifted to another BS in order to balance the network load.

Handover refers to the mechanism by which an ongoing session is transferred from one BS to another as seen in Fig. 2.2. Therefore, a *handover* decision mechanism is indispensable function of a cellular network. The decision for handover could be based on several parameters: signal strength, signal to interference ratio, distance to the base station, velocity, load, etc. The performance of the handover mechanism is extremely important in mobile cellular networks, in maintaining the desired quality of service (QoS).

For instance, Fig. 2.3 illustrates a typical signal strength reading as mobile station traverses to another cell. A handover decision may be triggered either when the target signal strength is higher than serving signal strength or when serving signal strength falls below a threshold. One can see that former may induce handover early but sustain better quality connection; however, the latter induces robust but poor quality connection since wireless channel introduces random large-scale variation in the received signal strength and handover decision mechanism based on measurements of signal strength induces the "ping-pong" effect, frequent handovers due to false triggers. Frequent handovers influence the QoS, increase the signaling overhead on the network, and degrade throughput in data communications. Thus,

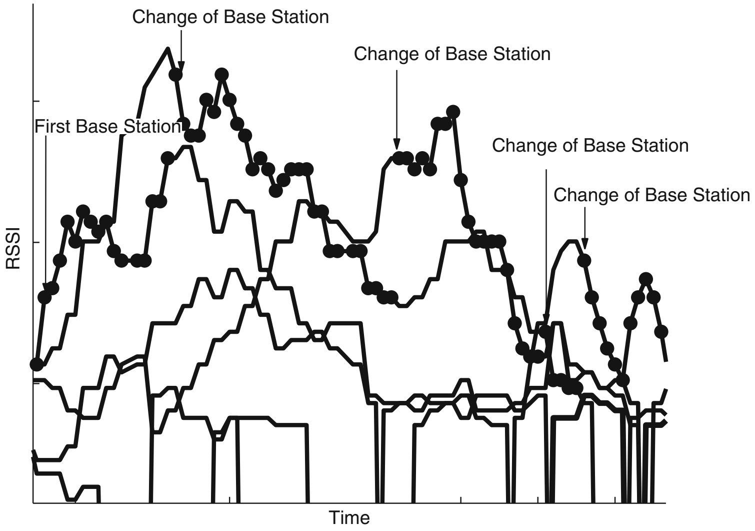


Fig. 2.2 RSSI readings and handover decision

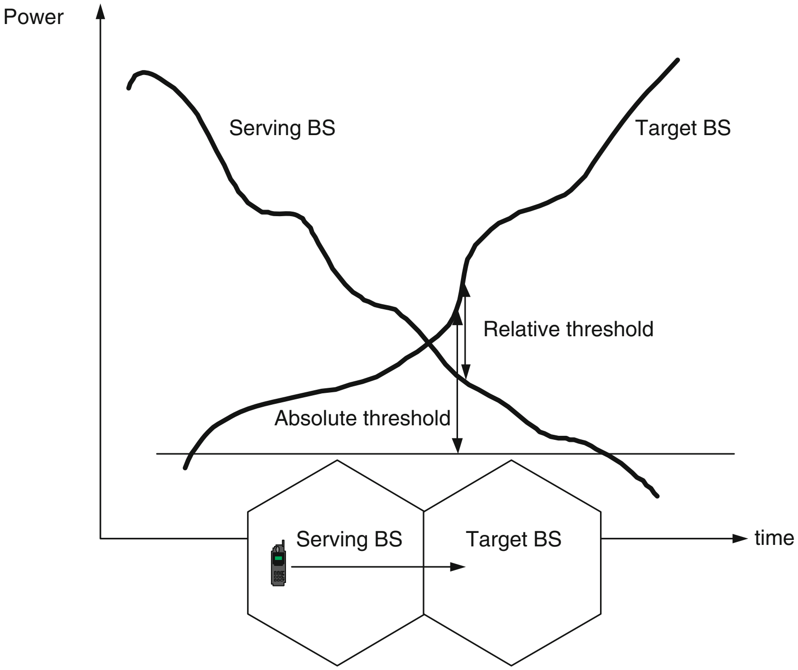


Fig. 2.3 HO decision

network operators should consider smart deployment strategies along with intelligent handover decision algorithms to efficiently use the network bandwidth while providing good connection.

2.1.2 Cellular Deployments

Capacity of cellular system is further being increased with advanced design techniques: cell splitting, sectorization, macro/micro cell, adaptive antennas, pico/femto cells, etc. These structures are depicted in Fig. 2.4.

Cell splitting is required when there is a demand for capacity more than the cell can offer. The cell is reduced to cover smaller area and number of cell sites are increased.

Also, to handle mobility better, *macro/micro cell* deployment is introduced. The wireless telecommunication system has a macro cell and at least one micro cell within the macro cell. Mobile stations with higher mobility are serviced by the macro cell and lower mobility mobile stations are handled with micro cells.

One cell can be further divided into multiple cells by *sectorization*. Sectorization uses sectoral antennas which have angle spread less than 360°. Sectorization

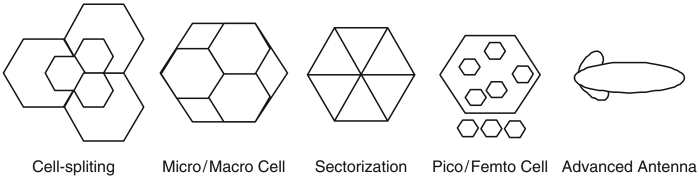


Fig. 2.4 Advanced Techniques to increase the capacity in cellular networks

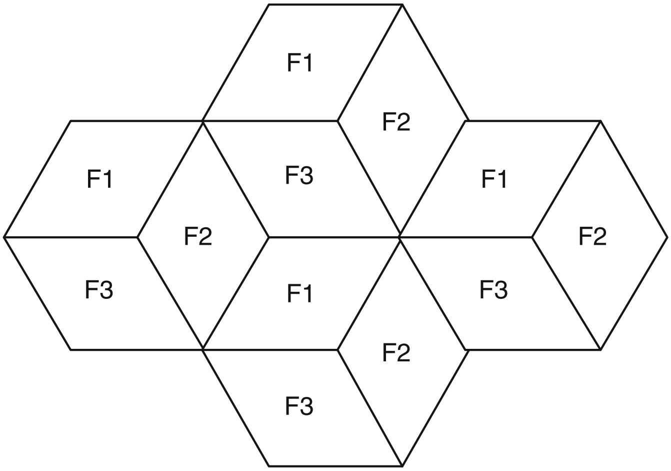


Fig. 2.5 Three sector deployment

increases the capacity with a factor of number of sector size. A typical deployment for three sector is shown in Fig. 2.5.

Lately, to further increase the indoor coverage, *picocells* and *femtocells* or "LCIB" (low-cost indoor/home base stations) are introduced, which are small coverage versions of the outdoor cellular base stations. Their connection to backbone is provided with an IP connection such as DSL or cable. These small cells are used to ensure in-building cellular coverage and may not require a macro BS. They are convenient but can cause challenges in the cell planning since they are bound to operate in the license bands and Carrier-to-Interference-plus-Noise Ratio (CINR) requirements for high speed data should be carefully maintained.

There are two approaches using picocells and femtocells: in-building for smaller sites and using an integrated picocell/distributed antenna system for midsize facilities. Femtocells have lower capacity than picocells and are designed for very small office spaces; however, picocells are used to cover buildings and streets. Additionally, picocell can be used with one radio and multiple spatially separated antennas, wired to the radio that resides in the pico BS. Distributed antennas can cover the building and relay the transmission to the pico BS. In this case, there will be no requirement for handover within the building.

Femtocell expands the coverage and provides better service to subscriber in terms of high speed, lower-latency, and lower battery consumption. Operator investment in femtocell is directly tied to subscriber demand when compared with investment in macro BS. Backhaul is over consumer broadband connection, which automatically decreases the operational expenses (OPEX). Operators may also provide the backhaul in some settings.

Another popular way of increasing efficiency is utilizing *multiantenna systems* at transmitter and/or receiver. Terms that are commonly associated with various aspects of multiantenna system technology include phased array, spatial division multiple access (SDMA), spatial processing, digital beamforming, adaptive antenna systems, and others.

Adaptive Antenna Systems (AAS) is one type of multiantenna system that introduces spatial processing systems with antenna arrays and signal processing modules. They adaptively change the radiation pattern of the radio environment to create spatially selective patterns. Spatially selective transmission increases the transmission rate and reduces the interference to nearby cells. They fall into two categories: switched-beam systems and adaptive array systems, as seen in Fig. 2.6. Switched beam antennas uses one of several predetermined fixed beams as the mobile station moves within the coverage of one base station. However, the most sophisticated adaptive array technology known as SDMA employs advanced signal processing techniques to locate and track mobile stations to steer the signals concurrently toward users and away from interferers. This directionality is achieved by at least 4–12 antenna elements.

Multiple-Input-Multiple-Output (**MIMO**) technology is another signal processing technique over multiantennas. MIMO promises to increase the capacity as well by creating independent channel with spatial separation. MIMO can be implemented standard off-the-shelf antennas as seen in Fig. 2.7. In cluttered environment, MIMO leverages multipath effects and works well; however, AAS beams become wider due to reflections.

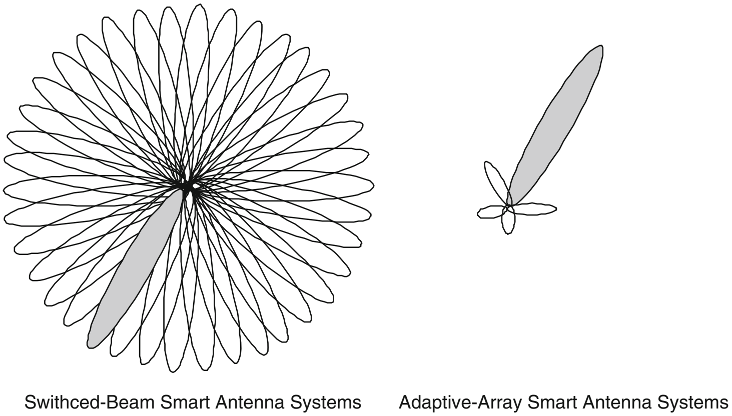
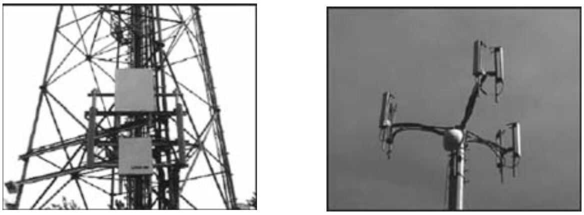

Swithced-Beam Smart Antenna Systems
Adaptive-Array Smart Antenna Systems

Fig. 2.6 Adaptive antenna systems



AAS

MIMO

Fig. 2.7 AAS and MIMO antennas

MIMO can offer more capacity by adding more antennas and more sectors. Capacity *linearly* increases with number of antennas by sending different information in different spatial streams. This is preferred if the channel is strong. However, if the signal is weak, MIMO sends the same information in different spatial streams to make the signal stronger. Unlike MIMO, AAS can only increase capacity by more powerful beam. But capacity pursue a *slow-logarithmic* growth with beam gain. For example, a WiMAX base station can offer 25 Mbps with one antenna in the transmitter and one antenna in the receiver (aka single-input-single-output -SISO-), a four-column AAS might increase this to 33 Mbps, and eight-column AAS can increase to just 38 Mbps. On the other hand, the capacity for 2×2 MIMO and 4×4 MIMO, is 50 Mbps and 100 Mbps, respectively. But, note that power consumption in the mobile station also increases with the number of antennas. We give more detail about MIMO in Chap. 6 as well as in Part 3 of the book.

2.2 Spectral Efficiency

Designing a cellular network trades off several competing requirements: capacity, service definition and quality, capital expenditures (CAPEX) and operational expenditures (OPEX), resource requirements including spectrum, end-user pricing/affordability, coexistence with other radio technologies. Lately, new developments such as femtocells and multiantenna system redefine this trade-off.

A metric called spectral efficiency is defined to quantify the efficiency of the cellular network. *Spectral efficiency* is a measure of the amount of information-billable services that carried by a cellular system per unit of spectrum. It is measured in *bits/second/Hz/cell*, which includes effects of multiple access method, digital communication methods, channel organization, and resource reuse.

To understand spectral efficiency calculations, consider the personal communications services (PCS) 1900 (GSM) system, which can be parameterized as follows: 200 KHz carriers, 8 time slots per carrier, 13.3 Kbps of user data per slot, effective reuse of 7 (i.e., effectively 7 channel groups at 100 percent network load, or only 1/7th of each channels throughput available per cell). The spectral efficiency is therefore: (8 slots × 13.3 Kbps/slot) / 200 KHz / 7 reuse = 0.08 b/s/Hz/cell.

Spectral efficiency is measured per cell meaning that the overall network efficiency is determined including the self generated interference. Thus, spectral efficiency is directly coupled to required amount of spectrum (CAPEX), required number of base stations (CAPEX, OPEX), required number of sites and associated site maintenance (OPEX) and ultimately, consumer pricing and affordability. The number of cells required is estimated by the following formula;

$$\text{number}-\text{of}-\text{cells/km}^2 = \frac{\text{offered}-\text{load}(\text{bits/s/km}^2)}{\text{available}-\text{spectrum}(\text{Hz})\times\text{spectral}-\text{efficiency}(\text{bits/s/Hz/cell})} \tag{2.1}$$

Note that there are three dimensions in the design: *spectral*, *temporal*, and *spatial*. We introduced the spatial tools in the previous section such as cellularization, sectorization, power control, multiple antennas, etc. Digital communication including modulation, channel coding, etc. and multiple access methods address the spectral and temporal components of the design. First, we introduce a summary of the digital communication in the next section and talk about the multiple access methods briefly in the subsequent section.

2.3 Digital Communication

Let us look at now the basics of digital communication that is being used in cellular networks. Digital communication is designed to transmit information sources to some destination in digital form whether the source is analog or digital. Analog source is converted to digital binary digits. It is originated in *telegraphy* era but modern digital communication has started in 1924 with *Nyquist sampling theorem*. The sampling theorem states that a signal bandlimited to **W** Hz can be reconstructed from the samples taken at **2W** pulses/s, which is the maximum pulse rate that can be achieved without any interference. Besides, this rate can be achieved with the following pulses $(\sin(2\pi Wt)/2\pi Wt)$ and analog source $x(t)$ is reconstructed with the following interpolation formula:

$$x(t) = \sum_{-\infty}^{\infty} x\left(\frac{n}{2W}\right) \frac{\sin[2\pi W(t-\frac{n}{2W})]}{2\pi W(t-\frac{n}{2W})}, \tag{2.2}$$

where $x(\frac{n}{2W})$ is the samples of $x(t)$ at Nyquist rate. Of course, samples are generally continuous. However, they are quantized into discrete values but with distortion.

Later, in 1948, Shannon introduced the mathematical foundation for information transmission in statistical terms. Shannon formula states that channel capacity is maximum mutual information of input and output as seen in Fig. 2.8;

$$C = \max_{p(x)} I(X;Y), \tag{2.3}$$

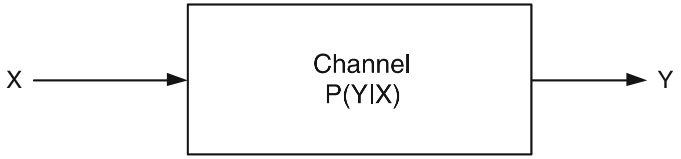


Fig. 2.8 Statistical channel; channel is conditional distributed given input

where mutual information $I(X;Y)$ is given by $I(X;Y) = h(Y) - h(Y|X)$ as the amount of information conveyed in the channel. In AWGN,[1] input and noise are independent. Consequently, the output is $Y = X + Z$, where $Z \sim N(0, N_o)$. Mutual information of input and output is given by $I(X,Y) = h(Y) - h(Z)$ since $h(Y|X) = h(Z|X) = h(Z)$. Let us restrict the input with constraint $E(X^2) \leq P$ and assume a Gaussian RV (random variable) input with variance P to maximize the capacity. Output Y is now a Gaussian RV with variance $N_o + P$ since they are independent. Capacity becomes

$$C = \frac{1}{2}\log_2(1 + \frac{P}{N}), \tag{2.4}$$

since the differential entropy of Gaussian RV with variance X is given by $\frac{1}{2}\log_2(2\pi eX)$. In general, a channel capacity that is bandlimited (W) is written as follows:

$$C = Wlog_2(1 + \frac{P}{WN_o}) = Wlog_2(1 + SNR)\text{bits/s}, \tag{2.5}$$

where WN_o is band-limited power spectral density of the additive noise and SNR is ratio of user's signal power to background noise, usually expressed in decibels (dB). For instance, for a communication channel with a bandwidth of 5 MHz and a signal to noise ratio of 20 dB, the channel capacity will be around 22 Mbps.

2.3.1 Source Coding

Shannon's statistical formula stemmed from the information measure introduced by Hartley in 1928. Hartley's formula stated that any information source produces a random output, which can be characterized statistically. As a result, number of possible choices from a finite set of equally likely outputs or any monotonic function of this number can be utilized as a measure of information.

Hartley introduced the logarithmic function as the measure since time, bandwidth, etc. tend to vary linearly with the logarithm of the number of possibilities. Therefore, (self-)information of the event $x = x_i$ is given by

$$I(x_i) = -\log_2(P(x_i)), \tag{2.6}$$

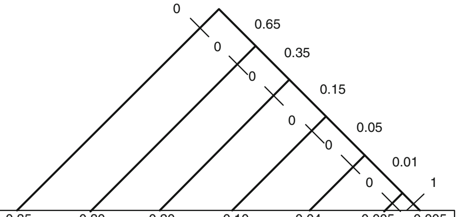
[1] Additive White Gaussian Noice.

where we can deduct that high probability event conveys less information than a low-probability event. Also, average self-information is denoted by *entropy* $H(x)$ as follows

$$H(x) = \sum_{i=1}^{m} P(x_i) I(x_i) = -\sum_{i=1}^{m} P(x_i) \log_2 P(x_i). \tag{2.7}$$

$H(x)$ is an important metric to find out the average number of binary digits required per output of the source. Typically, the channel has a lower bandwidth than the input signal bandwidth, and consequently the input signal has to be represented with less number of binary digits so that it can be accommodated on the channel. Reducing the redundancy in the information source is called source coding and entropy gives the shortest average message length (in bits) that can be sent to communicate the true value of the random variable to a recipient. Basically, it quantifies the valuable information in a piece of data. Of course, source is assumed to be memoryless, i.e., a source produces symbols that are statistically independent to each other. The discrete memoryless source (DMS) can be the simple model that can be used as a mathematical model.

Source coding may be over fixed or variable lengths; if the number of digits are fixed then it requires R ($=log_2 m$) bits when there are m symbols in the finite alphabet. Note that $R \geq H(x)$ and efficiency is defined as $H(x)/R$. In 1952, Huffman introduced a variable-length encoding algorithm for lossless data transmission. Huffman coding considers source alphabet probabilities $P(x_i)$, $i = 1, 2, \ldots, L$ and constructs a tree based on these probabilities as seen in Fig. 2.9. The probabilities of two least probable symbols are branched to compare with the next symbol. This is repeated until the high probable symbol. Note that Huffman coding requires knowing the probabilities of the symbols. Later, Lempel-Ziv (LZ[2]) algorithm is introduced 1977 and 1978 as an *universal source coding* algorithm, which does not require the source statistics where both compressor and decompressor create a dictionary on the fly.



Probability	0.35	0.30	0.20	0.10	0.04	0.005	0.005
Self-Information	1.5146	1.7370	2.3219	3.3219	4.6439	7.6439	7.6439
Code	**0**	**10**	**110**	**1110**	**11110**	**111110**	**111111**
Symbols	x1	x2	x3	x4	x5	x6	x7

H(X)=2.11
Average R =2.21

Fig. 2.9 A code for DMS

[2] This algorithm forms the basis for many LZ variations:

- LZ of 1977 (LZ77): LZR, LZSS, LZH, LZB, LZFG
- LZ of 1978 (LZ78): LZFG, LZC, LZT, LZMW, LZW, LZJ

2.3.2 Channel Coding

We introduced source coding as the procedure to represent the source information with the minimum number of bits. But when a code is transmitted over a noisy channel, error will occur. However, to achieve lossless transmission, additional redundancy needs to be introduced. As a result, the task of channel coding is to represent the source information in a manner that minimizes the error probability in the decoding. As a result, a channel code is longer than the source code so as to identify the correct input even if a few errors occur in transmission. Channel coding ensures that hamming distance between two codewords is larger than the resultant Hamming distance after transmission. For instance, codewords "100" and "011" have Hamming distance $d = 3$ since 3 bits disagree. Single bit error would change "100" to "000", "110" or "101". However, the closest would be still "100".

If we revisit the Shannon's capacity formula as illustrated in Fig. 2.10, channel coding theorem states that there exists a channel code that will permit the error-free transmission across the channel at a rate R, provided that $R \leq C$. Equality is achieved only when the *SNR* is infinite. There are two usages of channel coding: either for error detection or error correction.

2.3.3 Error Detection Coding

Error detection coding (EDC) is used only to detect errors. When error is detected, the receiver informs the transmitter for retransmission through ARQ (Automatic Repeat Request) mechanism. ARQ is an error control method, which uses acknowledgements and timeouts. ARQ resides in MAC or transport layer of the stack (OSI model). The most common EDC is *parity check coding*. This code only appends one bit to the end of m data bits to make $(m+1)$ bits even (or odd). A single bit error makes the even bit odd (odd bit even), consequently received code with error is separated by a Hamming distance of 2 or more. Parity checking is good protection against single and multiple bit errors when errors are independent.

More correlated errors that came in groups or bursts are handled with polynomial coding. Again, polynomial codes operate on the frame and append additional bits to the end of each frame. Typically, 16 or 32 bits are added and called either frame check sequence (FCS) or cyclic redundancy check (CRC).

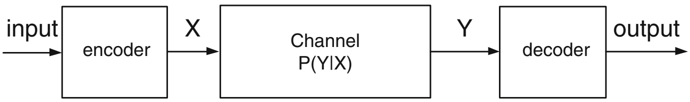


Fig. 2.10 Channel coding: Shannon capacity states that reliable information rate is possible but it does not specify how

In CRC, each frame is divided by a generator polynomial and the remainder of the division is added to the frame.[3] In the receiver, the division is repeated and results should be zero since remainder is already added in the transmitter. If division in the receiver is nonzero then error is detected. Notice that any error burst of length less than or equal to the length of the generator polynomial can be detected. The most commonly used polynomial lengths are as follows:

- 9 bits (CRC-8)
- 17 bits (CRC-16)
- 33 bits (CRC-32)
- 65 bits (CRC-64)

where CRC-16 and CRC-32 are widely used in IEEE 802.16e and in upper layer TCP/IP stack. In TCP/IP, stack error detection is performed at multiple levels. Ethernet frame carries a CRC-32 checksum. The IPv4 header contains another header checksum. Checksum is not used in IPv6 since it implements error detection. UDP has optional checksum and TCP has a checksum for payload. Packets with incorrect checksums are discarded and retransmission is requested.

2.3.4 Forward Error Correction

Error correction codes not only detect the error but also correct the errors to some extent. They are referred as Forward Error Correction (FEC) and can be classified into two types: block codes and convolutional codes. For example, Reed-Solomon (RS) coding, a block error correction coding, transforms a chunk of bits into a (longer) chunk of bits in such a way that errors up to some threshold in each block can be detected and corrected. More information is given about error correction coding in Chap. 4. However, we introduce decoding types in the following section.

2.3.5 Hard and Soft Decision Decoding

There can be two types of decoding: hard decoding and soft decoding. Hard decoding operates on the binary output of the demodulator. There is no side information such as the actual size of error in the analog domain. If analog error is used in the decoding operation better performance can be achieved. This is called "soft decoding," which is more complex and more optimal than hard decoding.

Introduction of *erasures* is another intermediate step in which decoder detects the coded bits and measures the reliability of the decision. If the reliability is low, the decoder outputs an erasure symbol, which is not a bit decision. As long as the Hamming distance d is equal to $2t+e+1$, t errors and e erasures can be corrected. This brings one half performance difference over hard and soft decision decoding.

[3] The parity bit coding, is in fact a CRC. It uses the two-bit-long divisor 11.

2.3.6 Puncturing

Note that error correction coding is defined by a *code rate*, which is defined as original length of the symbol over coded symbol length. For example, a code rate 1/2 introduces one additional bit per an input bit. To change the code rate of the encoded code, puncturing is introduced. On the one hand, puncturing further removes some parity bits to increase the coding rate. On the other hand, puncturing allows same low rate and low complexity decoder to be used for high rate encoded signal. Puncturing pattern sometimes shared by the receiver as well in order to do depuncturing. Higher coding ratios of 2/3 and 3/4 are obtained by puncturing the 1/2 rate code as follows; when 2 out of 6 bits are omitted the resulting rate becomes 3/4 and when 1 out of 4 bits is omitted the result gives a code with a rate 2/3.

2.3.7 Hybrid ARQ

To speed up the retransmission of frames received in error, ARQ mitigated from the MAC layer to the physical layer. Hybrid ARQ, a variation of ARQ, reduces the retransmissions with redundancy.

In ARQ, error detection bits (CRC) are considered to decide for retransmission. In HARQ, in addition to error detection bits, error correction bits are also added. This increases the robustness when the signal experiences bad channel. The two fundamental forms of Hybrid ARQ are chase combining (CC) and incremental redundancy (IR).

- Type I: *Chase combining* repeats the first transmission or part of it. A robust AMC can be achieved with chase combining. Chase combining suffers the capacity loss in strong signal condition. In terms of throughput, standard ARQ typically expends a few percent of channel capacity for reliable protection against error, while FEC ordinarily expends half or more of all channel capacity for channel improvement. Chase combining is used in HSDPA, Mobile WiMAX, and LTE.
- Type II: *Incremental redundancy* offers better performance with higher code rates in the beginning at the cost of additional memory and decoding complexity. IR does not have capacity loss in good signal, because FEC bits are only transmitted on subsequent retransmissions as needed. Information bits are encoded by a low rate mother code and family of high rate codes are obtained by puncturing the mother code as seen in Fig. 2.11. If a transmission is not successful, transmitter sends another higher rate code from the family. Consequently, each retransmission produces a codeword of a stronger code. In strong signal Type II Hybrid ARQ performs with as good capacity as standard ARQ. Incremental redundancy is used in 1xEVDO.

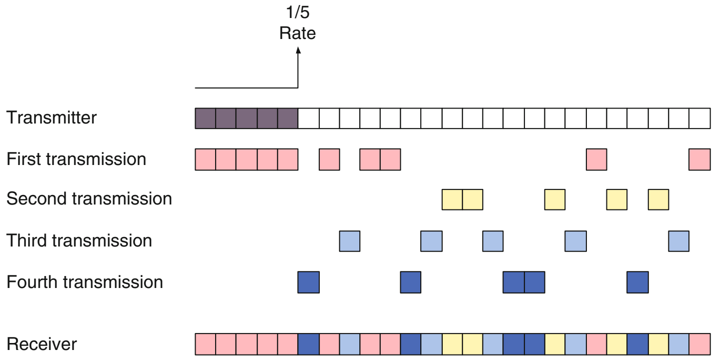


Fig. 2.11 HARQ Type II: Incremental redundancy

2.3.8 *Interleaving*

Another important component of digital communication is interleaving. Interleaving is used to combat for errors that occur in bursts. This is highly common in practice. Interleaving is basically used to shuffle the bits in the message after coding. Consequently, bursty errors are scattered when the bits are de-interleaved before being decoded. For instance, following gives an example of interleaving;

Error-free transmission	mmmmuuuusssstttteeeerrrrgggg
Transmission with a burst error	mmmmuuuusss____teeeerrrrgggg
Interleaved	mustergmustergmustergmusterg
Interleaved Transmission with a burst error	mustergmust____ustergmusterg
After deinterleaving	mm_muuuusssstttte_eer_rrg_gg

We can see that in each codeword {mmmm, eeee, rrrr, gggg}, only one bit is altered. As a result, one bit error correcting code is ample to correct everything correctly. Of course, latency is increased since in order to decode the first codeword all the codewords have to be received.

2.3.9 *Encryption and Authentication*

Another level of manipulation on the information before transmission is applying an *encryption* scheme, which is necessary to protect bits over the channel for various kind of attacks. Also, sometimes, message only needs to be *authenticated* against unprotected altering during transmission.

A cipher is used to encrypt the plaintext and decrypt the ciphertext, which is not understandable without decrypting it. The ciphers are either block ciphers or continuous stream ciphers. Block ciphers de/encrypts the fixed size plaintext in contiguous blocks; however, stream ciphers de/encrypts a continuous stream of symbols.

The cipher output depends on a *key*, which changes the ciphertext as compared to plaintext. If the same key is used for encryption and decryption then it is a symmetric key algorithm, otherwise asymmetric key algorithm.[4]

The Advanced Encryption Standard (AES) is one of the widely used symmetric key cryptography. AES, introduced in 1997 by US National Institute of Standards and Technology (NIST), is the successor of Data Encryption Standard (DES), which was found too weak because of its small key size and the technological advancements in processor power. The Rijndael, whose name is based on the names of its two inventors from Belgium, Joan Daemen and Vincent Rijmen, is the version introduced in 2000 after a contest.

The Rijndael is a block cipher, which takes an input block of a certain size, usually 128, and produces a corresponding output block of the same size. The transformation requires a second input, which is the secret key. It is important to know that the secret key can be of any size (depending on the cipher used) and that AES[5] uses three different key sizes: 128, 192, and 256 bits. Each block constitutes a 4 × 4 *state* matrix and there is a *key* matrix, which is also 4 × 4 in size. The Rijndael algorithm defines the following steps as seen in Fig. 2.12:

- *SubBytes:* Each byte is replaced with another according to a lookup table.
- *ShiftRows:* Each row of the state is shifted cyclically by a certain number
- *MixColumns:* Each column of the state is considered as an input to produce an four bytes output as a new column.
- *AddRoundKey:* Each byte is combined with a subkey, which is derived from the main key using Rijndael's key schedule algorithm.

Encryption takes couple of rounds and number of rounds differ depending on the key sizes, for a key size of 128, it requires 10 rounds; however, for key size 192 (256), it requires 12 (14) rounds. Note that in the final round there is no MixColumns step as seen in Table 2.1.

Authentication is used to detect whether the message is corrupted or not. Previously, we introduce CRC, which is used to detect errors in a message. Similar to CRC, a *message digest* that is calculated by using a hash function to act as an unique fingerprint of the message is used to detect the corrupted messages. It provides stronger assurance of data integrity than check sum. However, message digest does not protect against unauthorized modification of the message since a forger

[4] "If the algorithm is symmetric, the key must be known to the recipient and to no one else. If the algorithm is an asymmetric one, the enciphering key is different from, but closely related to, the deciphering key. If one key cannot be deduced from the other, the asymmetric key algorithm has the public/private key property and one of the keys may be made public without loss of confidentiality...".

[5] "While AES supports only block sizes of 128 bits and key sizes of 128, 192, and 256 bits, the original Rijndael supports key and block sizes in any multiple of 32, with a minimum of 128 and a maximum of 256 bits."

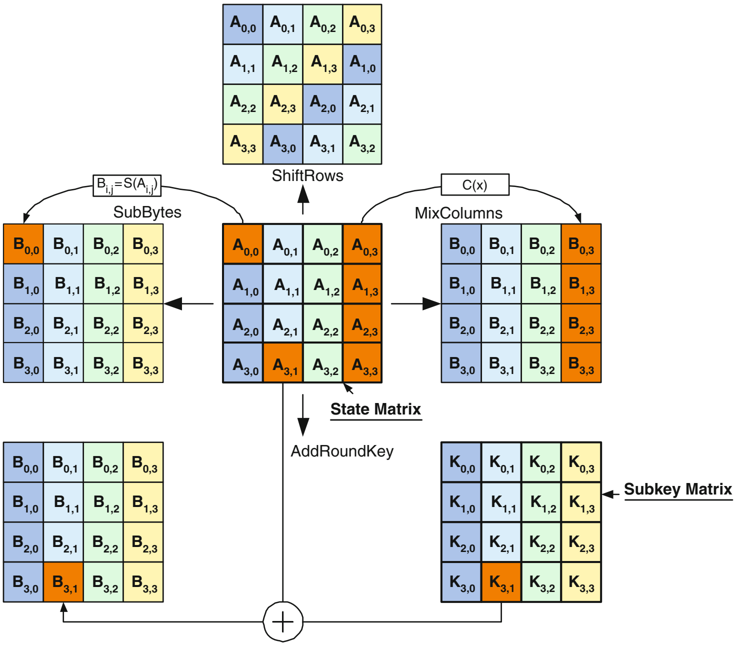

$A_{0,0}$ $A_{0,1}$ $A_{0,2}$ $A_{0,3}$
$A_{1,1}$ $A_{1,2}$ $A_{1,3}$ $A_{1,0}$
$A_{2,2}$ $A_{2,3}$ $A_{2,0}$ $A_{2,1}$
$A_{3,3}$ $A_{3,0}$ $A_{3,1}$ $A_{3,2}$
ShiftRows
$B_{i,j}=S(A_{i,j})$
SubBytes
C(x)
MixColumns
State Matrix
AddRoundKey
Subkey Matrix
$B_{0,0}$ $B_{0,1}$ $B_{0,2}$ $B_{0,3}$
$K_{0,0}$ $K_{0,1}$ $K_{0,2}$ $K_{0,3}$

Fig. 2.12 AES

Table 2.1 The AES rounds

Initial round	AddRoundKey(state, roundkey[round])

Rounds	SubBytes(state)
	ShiftRows(state)
	MixColumns(state)
	AddRoundKey(state, roundkey[round])

Final round	SubBytes(state)
	ShiftRows(state)
	AddRoundKey(state, roundkey[round+1])

Output	state

can create an alternative message and its corresponding message digest (MD5 or SHA hash value). To protect the message against this type of attack, a secret key can also be used during the hash. This only allows the owner of the secret key to produce the valid message digest, which is now called a Message Authentication Code (MAC). It is a symmetric key solution since shared secret key needs to be known at the transmitter and the receiver. If we call this Hash-MAC (HMAC) [RFC2104, RFC2202], then there is another MAC algorithm, called Cipher based MAC (CMAC) or AES-CMAC [RFC4493, RFC4494, RFC4615] since it is a keyed hash function based on symmetric key block cipher, such as the AES. AES-CMAC is a variation of CBC-MAC (Cipher Block Chaining Message Authentication Code) and responsible to generate a MAC (T), a 128-bit string with three inputs: key, message (M) and message length (len) as follows; T = *AES-CMAC*(key,M,len). CBC-MAC is responsible to create blocks out of M as $m_1, m_2, \cdots, m_k$ and encrypts block i with a block cipher that takes the block i, key and encrypted output of block $i-1$. Note that *initialization vector* for block 1 is 0.

If a message is protected with MAC, it can only be forged by creating a second document that has the same hash value as the original document. The forged document may contain a different text, which are altered repeatedly until the computed hash value matches. However, this needs 2^m (2^{128} trials for MD5) trials if the hash value is m bits. It is almost impossible in a timely manner.

The IEEE 802.16e (WiMAX-e) utilizes AES in CCM (Counter with CBC-MAC) mode. The CCM mode combines the counter (CTR) mode of encryption with the CBC-MAC mode of authentication and uses 128-bit key AES. Counter mode generates blocks of 128bits in size from the message and XOR them with a block of the key stream, which is generated by the AES encryption of an arbitrary value. Arbitrary value is called the *counter* and it generally differs by 1 between each adjacent blocks. AES-CTR is selected since it provides a convenient way for decryption as well.

2.3.10 Digital Modulation

Finally, digital bit stream is converted to analog waveform for Radio Frequency (RF) bandpass channel. Quadrature amplitude modulation (QAM) is one of the widely used modulation scheme, which changes (modulates) the amplitude of two sinusoidal out of phase carrier waves depending on the input bits as follows:

$$s(t) = I(t)\cos(2\pi f_c t) + Q(t)\sin(2\pi f_c t), \tag{2.8}$$

where $I(t)$ and $Q(t)$ are the modulating signals from the QAM constellation and f_{c} is the carrier frequency.

In QAM, the constellation points are usually arranged in a square grid with equal vertical and horizontal spacing as seen in Fig. 2.13; however, other configurations are also possible (e.g., Cross-QAM). The most common forms of QAM constellation are 16QAM, 64QAM, 128-QAM, and 256-QAM. Higher-level constellation enables transmitting more bits per symbol; however, if the mean energy is kept constant, then the points in the constellation comes closer as constellation size gets higher. Of course, it now becomes more susceptible to error. As a result, higher-order QAM can deliver more data less reliably than lower-order QAM.

Figure 2.14 shows the ideal constellation diagram as well as the constellation diagrams with impairments. Any imperfection in the transmission and receiving process may cause shifts of the points in the constellation, which if large may result in an error in the demodulation process. Now, we introduce below some of the imperfections that are observed commonly.

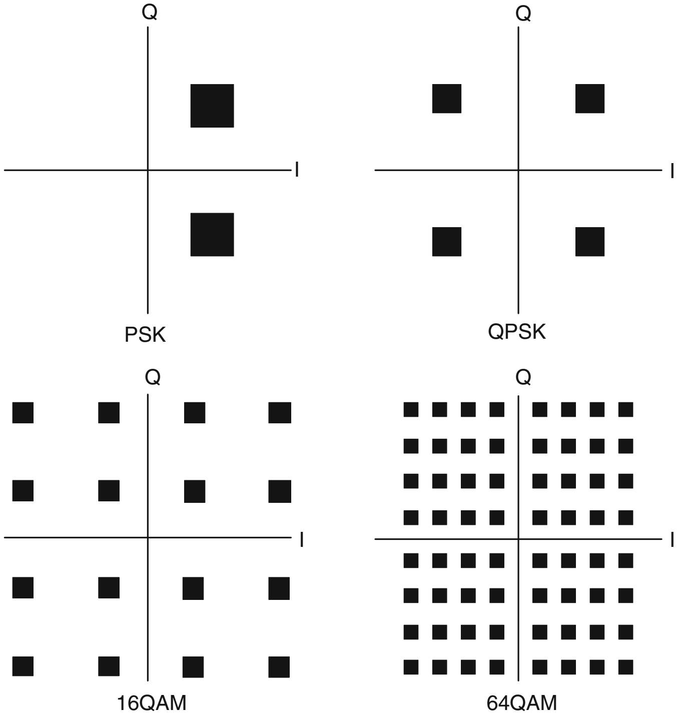


Fig. 2.13 QAM constellation diagrams

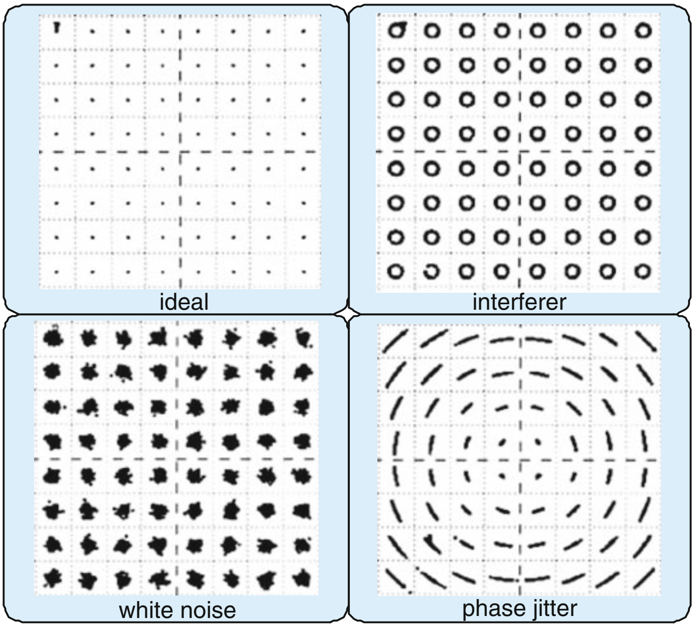


Fig. 2.14 QAM imperfections. Source: http://www.blondertongue.com

- **Amplitude Imbalance** describes the different gains of the I and Q components of a signal. In a constellation diagram, amplitude imbalance shows by one signal component being expanded and the other one being compressed. This is because the receiver AGC (Automatic Gain Controller) makes a constant average signal level.
- **Phase Error** is the difference between the phase angles of the I and Q components referred to 90. A phase error is caused by an error of the phase shifter of the I/Q modulator. The I and Q components in this case are not orthogonal to each other after demodulation.
- **Interferers** are sinusoidal signals that operate in the same frequency range and superimpose on the QAM signal at some point in the transmission path. An interferer is shown in the constellation as a rotating pointer as seen in Fig. 2.14. Path of pointer constructs a circle around each ideal signal point. For example, if there is a leakage from an interferer with the same frequency then superimposed constellation occurs.
- **Additive Gaussian noise** has an additive effect, which superimposes on the constellation point. Normally, it has a constant power density and a Gaussian amplitude distribution throughout the bandwidth of a channel. It appears as clouds around the constellation points as seen in Fig. 2.14.
- **Phase Jitter** in the QAM signal is caused by transmitter in the transmission path or by the I/Q modulator. In contrast to the phase error, phase jitter is a statistical quantity that affects the I and Q path equally. Phase jitter causes signal states being shifted about their coordinate origin as seen in Fig. 2.14.

2.4 Wireless Channel

Before we offer arguments for why OFDM is selected as the transport technology of the next generation wireless communication, let us first introduce you what wireless channel looks like. For your first look at wireless we are going to talk about the statistical characterization of a wireless channel, which will be a reference for many chapters in the book. The goal here is to give you a sense of how the communication is affected by the wireless channel and what the parameters are that needs to be taken into account during the design. Of course, wireless channel is a topic of its own and in the bibliography section, we list some of the good books that details. In the interest of brevity, we are going to be a little terse but intuitive – this is just an introduction, but we hope you will find it fun and useful.

After digital modulation, the constructed signal is sent to antenna to be transmitted over the air. When a voltage is applied to an antenna, it creates electromagnetic field that propagates according to Maxwell's equation[6] that states that

[6] "The mathematical expression for the E_z vertical component of a random electric field is given by

$$E_z(r,\varphi) = E_0 \sum_{-N}^{N} i^n \alpha_n J_n\left(\frac{2\pi}{\lambda_w} r\right) e^{in\varphi}, \tag{2.9}$$

electromagnetic fields propagate in free space in all directions (note that this is affected by antenna geometry which determines how power flows in any given direction), like light. These waves induce electric currents in the receiver's antenna but the energy it creates for a given voltage of a given frequency is directly coupled with antenna size. Antenna size is directly coupled with the field's wavelength (λ), which is basically inversely proportional with the operating frequency (f_c) as follows; $\lambda = c/f_c$ where c is speed of light. As a result, higher the frequency, the smaller the antenna size. Moreover, antennas could be in different characteristics, omni-directional antennas boosts the power in all directions, while sectoral ones just empowers in certain directions.

Signals travel at a finite speed (upper limit is speed of light), a receiver senses a transmitted signal only after a time delay (Δt) directly related to the propagation speed and the distance (d): $\Delta t = d/c$. For instance, 16 ms is enough to travel across the United States at the speed of light and at least three times longer if transmitted through a coaxial cable, which connects the East and West coasts. However, the signal strength diminishes as it traverses further away from the transmitter. This is due to the conservation of energy principle since amount of energy given to the free space should be constant and independent from the distance d. The total power can be found by integrating a sphere centered at the transmitter as follows; $P(d)4\pi d^2$, which is the total energy at distance d then $P(d)$ should be inversely proportional to d^2. This makes received signal amplitude V_R proportional to transmitted signal amplitude V_T as follows; $V_R = |KA_T/d|$ for some constant K but with a phase shift of $\mathrm{e}^{-j2\pi d/\lambda}$. Note that this is the number one enabler of the cellularization since a signal almost disappears as it traverse further away from the transmitter. This phenomenon is called *pathloss* effect, which is one of the large-scale characteristics of the wireless channel. There are also *shadowing* and *fading* effects that stand for medium and short-term characteristics of the wireless channel.

2.4.1 Pathloss

Earlier, we showed the underlying mechanism of free-space model that is used to characterize the pathloss when there is no extra attenuation due to physical characteristic of the environment. However, if the physical environment is open to reflection or scattering or diffraction (aka multipath components) as seen in Fig. 2.15, then Maxwell's propagation equation can be extended to formulate pathloss via *ray-tracing*. However, ray-tracing only produces accurate results if multipath components are small. if otherwise, empirical models are present to provide accurate formulation based on channel measurements.

where E_z is given for two-dimensional omni-directional wireless channel and E_0 denotes the mean-square value of the electric field vertical component; r and φ designate the polar coordinates in the center wavelength, and α_n represents statistically independent complex random values with zero mean and unit variance. The $(2N+1)$ number of coefficients α_n defines the size of the area to which the model applies".

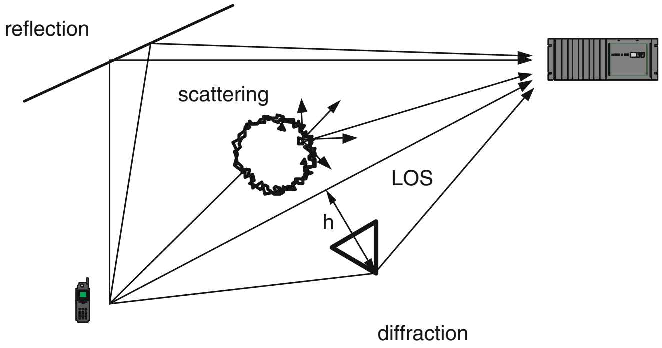


Fig. 2.15 Multipath components

Let us first introduce the free-space path loss model as follows;

$$\frac{P_{\mathrm{r}}}{P_{\mathrm{t}}} = \left[\frac{\sqrt{G_l}\lambda}{4\pi d}\right]^2, \tag{2.10}$$

where P_{r} and P_{t} are received and transmitted power, respectively. Note that this formulation is for line-of-sight (LOS) communication in which there is only one direct path from transmitter to the receiver. Therefore, product of antenna field ($\sqrt{G_l}$) at the corresponding peers is important in determining the reception power.

This formula, however, is not ample to formulate the pathloss if multipath components are present. Ray tracing is a way to offer path loss formulation by assuming finite number of reflectors since a reflector is considered as another source and the resultant signal at the receiver is the sum of signals coming from all sources in different distances and altered antenna fields.

Two-ray model is the simplest ray tracing model and only considers a LOS component and a reflected path from the ground as seen in Fig. 2.16. We skip the details in the derivation and only present the approximated two-ray model for large d and equal antenna field ($G_l = G_i$) as follows;

$$\frac{P_{\mathrm{r}}}{P_{\mathrm{t}}} = \left[\frac{\sqrt{G_{\mathrm{l}}}h_{\mathrm{t}}h_{\mathrm{r}}}{d^2}\right]^2, \tag{2.11}$$

where d equals to $x+y$ as well as l consequently δ is approximated as zero. This equation shows that unlike free space the signal attenuates at large distances with $\approx 1/d^4$. One can see here the single reflection from ground may attenuate the signal severely at large distances. This phenomenon is explained as follows:

- If $d < h_{\mathrm{t}}$, signals add up constructively and path loss is proportional to $1/(d^2 + h_{\mathrm{t}}^2)$ for $h_{\mathrm{t}} \gg h_{\mathrm{r}}$.

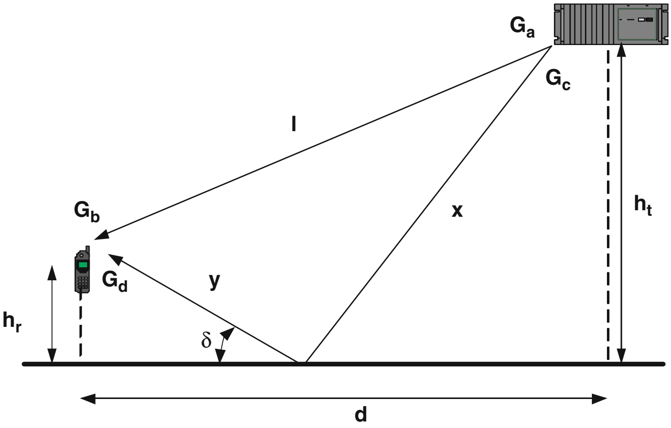


Fig. 2.16 Two-ray module

- If $h_{\mathrm{t}} < d < d_c$, signals experience constructive and destructive interference up to a critical distance $d_c \approx 4h_{\mathrm{t}}h_{\mathrm{r}}/\lambda$.
- If $d_{\mathrm{c}} < d$, signals experience destructive interference and path loss is proportional to $1/d^4$.

There are several more sophisticated ray tracing models that consider more than one reflection. Moreover, general ray tracing formula also considers rays from diffraction and rays from scattering as well.[7]

2.4.1.1 General Pathloss Formula

In general, following formulation is used for pathloss (PL)

$$\begin{aligned} P_{\mathrm{r}} &= P_{\mathrm{t}}K\left[\tfrac{d_o}{d}\right]^{\gamma}, \\ PL &= K\,dBm - 10\gamma\log_{10}\left[\tfrac{d_o}{d}\right], \end{aligned} \tag{2.12}$$

where $PL = P_{\mathrm{r}} - P_{\mathrm{t}}$ in dBm and d_o is reference distance for the antenna far field. K is unitless constant given in dBm by the following formula;

$$K\,dB = 20\log_{10}\frac{\lambda}{4\pi d_o}. \tag{2.13}$$

γ is the pathloss exponent and typically varies for different environments; γ ranges between 3.7 and 6.5 for urban macrocells and 2.7 and 3.5 for urban

[7] "Diffraction loss is commonly modeled with Fresnel knife-edge diffraction, which formulates the delay and path loss. Scattering also introduces additional delay and alters the antenna gain according to geometry of the scattering object".

microcells; for indoor γ falls below 3.5; however, it ranges between 2 and 6 for multiple floor office spaces. The values for γ are obtained from the empirical methods described next.

We put more emphasis on empirical models since typically mobile communication is in complex propagation environment and can only be approximated with real data from the field. There are numerous pathloss models introduced for various configuration and setting that uses power measurements. Measurements from the field are averaged over time and wavelength to filter off the multipath effects. Also, measurements of multiple locations with same characteristics are averaged to get a general formulation for the respective characteristic (e.g., urban macro cell).

2.4.1.2 Hata-Okumura Model

In 1968, Okumura has conducted measurements of base station to mobile station in Tokyo and introduced empirical plots. Later, in 1980, Hata developed a closed-form expression from Okumura's data. Hata-Okumura model is the most widely used path loss model for macrocellular environments. It is valid for the 500–1500 MHz frequency range. Receiver distance is greater than 1km from the base station where base station antenna heights are greater than 30 m. The analytical approach to the model is given in dB as follows;

$$\begin{aligned} PL(\text{urban}) = 69.55 + 26.16\log_{10}(f_c) - 13.82\log_{10}(h_\text{t}) \\ - C_\text{H} + [44.9 - 6.55\log_{10}(h_\text{t})]\log_{10}(d), \end{aligned} \tag{2.14}$$

where C_H is the antenna height correction and given as follows for small or medium sized city,

$$C_\text{H} = 0.8 + (1.1\log_{10} f_c - 0.7)h_\text{m} - 1.56\log_{10} f_c, \tag{2.15}$$

and for large cities,

$$C_\text{H} = \begin{cases} 8.29(\log_{10}(1.54h_\text{m}))^2 - 1.1, & 150 \leq f_c \leq 200 \\ 3.2(\log_{10}(1.75h_\text{m}))^2 - 4.97, & 200 \leq f_c \leq 1500 \end{cases} \tag{2.16}$$

h_m ($=h_\text{r}$) is the height of the mobile station. For suburban areas it is modified as follows;

$$PL(\text{suburban}) = P(\text{urban}) - 2\left(\log_{10}\frac{f_c}{28}\right)^2 - 5.4\ . \tag{2.17}$$

Hata-Okumura model is developed for large cells and BS is assumed to be higher than the rooftops. These models are developed for first generation systems and may not work well with WiMAX and 4G that deploys smaller cell size operating with higher frequencies.

2.4.1.3 COST-231 Walfish-Ikegami (W-I) Model

The European Cooperative for Scientific and Technical (COST) research extended the Hata-Okumura model for large and macro cells to 2 GHz as follows

$$PL = 46.3 + 33.9\log_{10} f_c - 13.82\log_{10} h_\mathrm{t} - C_\mathrm{H} + [44.9 - 6.55\log_{10} h_\mathrm{t}]\log_{10} d + C, \tag{2.18}$$

where C equals to 0 dB for medium cities and suburban areas and 3 dB for metropolitan areas. C_H is the antenna height correction of the Hata-Okumura model. However, this model is restricted to frequencies between 1.5 and 2 GHz with base station height around 30–300 m and mobile height around 1–10 m.

Later, COST-231 group has combined the findings proposed by Walfish-Ikegami (W-I)[8] to introduce model for micro and small macro cells (d is between 0.02 and 5 km). It extends the Hata-Okumura model for flat suburban and urban areas with uniform building height. It is applicable for frequencies between 800 and 2,000 MHz with base station height ranging from 4 to 50 m. COST-231 W-I model introduces two formulation for LOS and NLOS cases; For LOS, pathloss is given by

$$PL = 42.6 + 26\log_{10}(d) + 20\log_{10}(f_c), \tag{2.19}$$

for $d \geq 0.02$ km. For NLOS case, pathloss consists of free-space pathloss (L_o), multi-screen loss (L_msd), and (L_rts), which is the loss from the last rooftop to the mobile station. Hence, pathloss is given as follows;

$$PL = L_\mathrm{o} + max(0, L_\mathrm{rts} + L_\mathrm{msd}), \tag{2.20}$$

where L_o is given as

$$L_\mathrm{o} = 32.4 + 20\log_{10} d + 20\log_{10} f_c. \tag{2.21}$$

We skip the details of L_rts and L_msd but briefly introduce them as follows; L_rts requires the width of the street as well as the difference between the building height and height of the mobile station. There is also correction factor which takes the street orientation in perspective. L_msd requires the difference between height of the base station and roof top level. COST-231 W-I is accepted to ITU-R (Report 567-4); however, the model does not give good performance if antenna heights are less than the rooftop level.

2.4.1.4 Erceg Model

As we said, Hata-Okumura model was found that it is not suitable for shorter base station heights. Erceg model used the experimental data collected by AT&T Wireless Services across United States in 95 existing macro cells operating at 1.9 GHz.

[8] Also called the Hata Model PCS Extension.

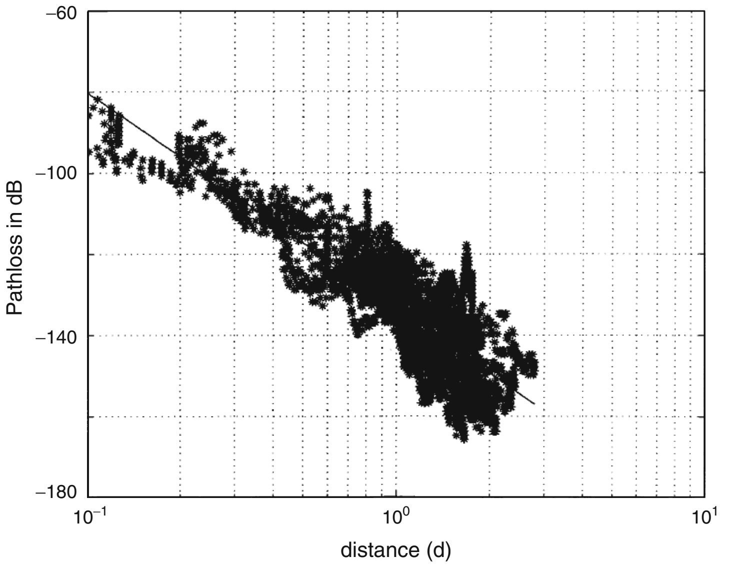


Fig. 2.17 Pathloss for a macrocell in the Seattle area: base station height is 25 m. Source: Erceg, IEEE JSAC, 1999

Scatter plots obtained in Seattle area is illustrated in Fig. 2.17. One can see the straight line representing the least-squares linear regression fit. Pathloss model introduced is written as

$$PL = A + 10\gamma \log_{10}\frac{d}{d_o} + \chi; \quad d \geq d_o, \tag{2.22}$$

where d_o is selected as 100 m and A is found to be close to the free-space formula and given by the following formula

$$A = 20\log_{10}(4\pi d_o/\lambda), \tag{2.23}$$

and path loss exponent γ (found to be greater than two) is expressed as

$$\gamma = (a - bh_t + c/h_t) + x\sigma_\gamma, \quad 10m \geq h_t \geq 80m, \tag{2.24}$$

where h_t is the base station height; σ_γ is the standard deviation of γ; x is a zero mean Gaussian variable of unit standard deviation, $N[0,1]$; and a, b, c and σ_γ are from the Table 2.2. The shadow fading component χ is a zero-mean Gaussian variable

$$\chi = y(\mu_\sigma + z\sigma_\sigma), \tag{2.25}$$

where y and z are zero-mean Gaussian variable with standard deviation $N[0,1]$. Typical standard deviation for χ is found to be between 8.2 and 10.6 dB, depending on

Table 2.2 Numerical values of model parameters. Source: Erceg, IEEE JSAC, 1999

Model parameter	Terrain A (hilly/ moderate-to-heavy tree density)	Terrain B (hilly/light tree density or flat/ moderate-to-heavy tree density)	Terrain C (flat/ light tree density)
a	4.6	4.0	3.6
b	0.0075	0.0065	0.0050
c	12.6	17.1	20.0
σ_γ	.57	.75	.59
μ_σ	10.6	9.6	8.2
σ_σ	2.3	3.0	1.6

the terrain. Note that correction terms are needed for different frequencies (ΔPL_f) and for receive antenna heights above 2 m (ΔPL_h). Also, Terrain C is a good match with the COST-231 W-I model for suburban areas as well.

2.4.2 Shadowing

As we introduced above, signal attenuation is a random process due to the multipath components. However, the random variations may show mid-term or short-term changes. Mid-term changes typically arise from the blocking objects between the transmitter and receiver and considered as *slow-fading* or *log-normal shadowing*. Fading is slow and predictable in the sense that shadowed areas tend to be large and blocking objects does not change their location, size, dielectric properties rapidly. However, it is the most severe attenuation factor and predominantly present in heavily built up areas. In brief, power attenuation with log-normal shadowing considers a log normal random variable as another attenuation factor in addition to the attenuation due to pathloss. The following formula combines the shadowing with pathloss formula as follows;

$$PL = K_{dBm} - 10\gamma \log_{10}\left[\frac{d_o}{d}\right] - \chi_{dBm}, \tag{2.26}$$

where χ is a Gaussian random variable with zero mean and variance σ_χ.

As we said, shadowing is detrimental but if it blocks the interference then it is beneficial. It is also interesting to note that radio signals behind hills avoid total shadowing due to the diffraction since signals bend over. Typically, severity is minimized by placing the antennas far apart to clear off more obstruction. Figure 2.18 shows an analysis of pathloss and shadowing with respect to different type of terminals. One can see the affect of line-of-sight and building loss.

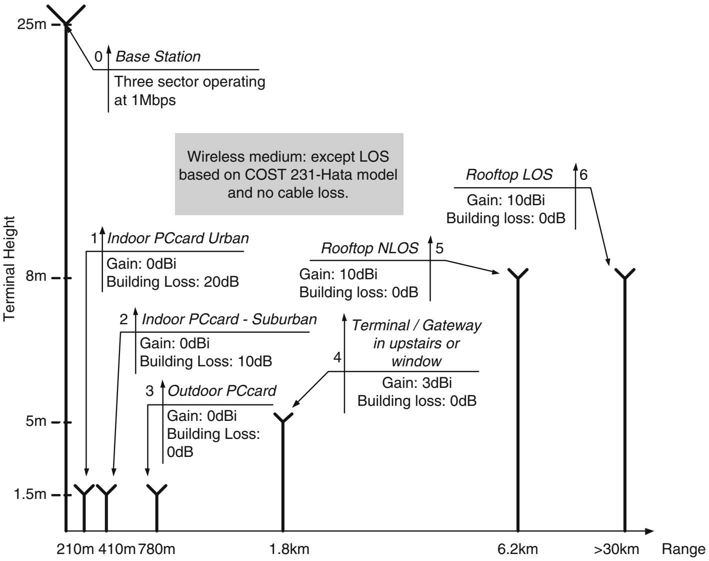


Fig. 2.18 Coverage analysis with respect to type of terminals. Source: Vodafone

2.4.3 Fading

Now, we start introducing the short-term variation in the signal due to multipath components. Previously, we introduced that there can be multipath components of a signal in addition to the LOS component and introduced the signal attenuation for deterministic channels. However, in real life, channel is *time-varying*. We analyze the statistical properties of the time-varying fading channels with *the Bello functions*, which was worked out by Bello in 1963. The first function ($h(\tau,t)$), as illustrated in Fig. 2.19, is the time-variant impulse response or input delay spread function.

$h(\tau,t)$ arises from the *pulse train* concept as illustrated in Fig. 2.20. A single pulse creates several *rays* where each ray either corresponds to LOS or multipath component. However, it is time-varying meaning that with time amplitude (α_i), phase ϕ_i, and delay (τ_i) of the ray changes. We can formulate the multipath channel $h(\tau,t)$ as follows;

$$h(\tau,t) = \sum_{n=0}^{N(t)} \alpha_n(t)\mathrm{e}^{-j\phi_n(t)}\delta(\tau - \tau_n(t)), \tag{2.27}$$

where $N(t)$ is the number of multipath rays and $\phi_n(t)$ is composed of phase shifts due to delays ($\tau_n(t)$) and *Doppler phase shift* ($\phi_{D_n}(t)$) as follows; $\phi_n(t) = 2\pi f_c\tau_n(t) - \phi_{D_n}(t)$. $2\pi f_c\tau_n(t)$ is the phase shifts due to multipath components and typically $f_c\tau_n(t) \gg 1$ since f_c is typically greater than 1GHz and $\tau_n(t)$ is on the

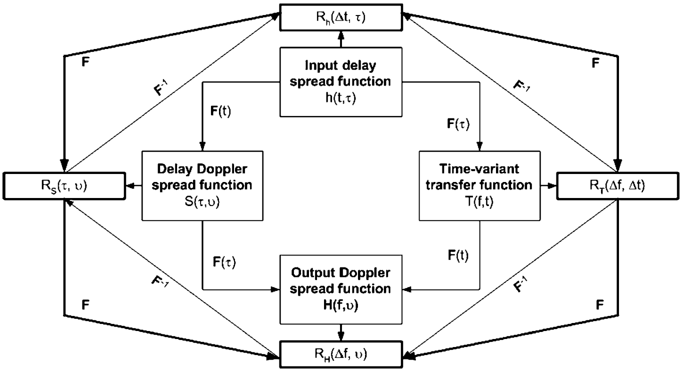

R_h(Δt, τ)
Input delay spread function h(t,τ)
F
F^-1
F(t)
F(τ)
F
F^-1
R_S(τ, υ)
Delay Doppler spread function S(τ,υ)
Time-variant transfer function T(f,t)
R_T(Δf, Δt)
F(τ)
F(t)
F^-1
F
Output Doppler spread function H(f,υ)
F^-1
F
R_H(Δf, υ)

Fig. 2.19 Bello functions

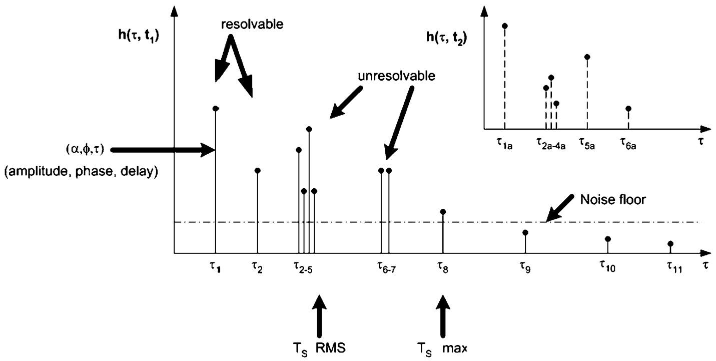


Fig. 2.20 Multipath time varying channel and delay spread: note that typically multipath components below the noise threshold is ignored. Typical delay spread in suburbs are around $0.2-20\,\mu s$; in urban environment it is around $1-30\,\mu s$ and for indoor environment it is $40-200$ ns

order of $1-20$ µs for outdoor. As a result, a small change in the delay may cause significant phase change, which may result in constructive or destructive effect in the signal.

Now, we introduce Doppler shift (f_D) first in order to describe $\phi_{D_n}(t)$. Doppler shift arises due to mobility since a mobility in the source or destination or any other objects in between may change the length ($r_n(t)$) of the ray by Δd. This change in distance depends on the directional velocity ($v(t)\cos\theta$) as follows; $\Delta d = v(t)\cos\theta\Delta t$, where θ is the angle between the LOS path and direction of motion. As a result, this results in a phase change ($\Delta\phi$) as follows;

$$\Delta\phi(t) = 2\pi f_\mathrm{c}\Delta d/c = 2\pi f_\mathrm{D}(t)\Delta t, \tag{2.28}$$

where f_c is the carrier frequency and c is the speed of light. This can be formulated by a shift in the carrier frequency $f_\mathrm{D}(t)$ as follows;

$$f_\mathrm{D}(t) = \frac{v(t)}{\lambda}\cos\theta, \tag{2.29}$$

where λ is the signal wavelength. We can further simplify by assuming that Doppler frequency is changing slowly within the time of interest: $f_{\mathrm{D}}(t) \approx f_{\mathrm{D}}$. As a result, Doppler phase shift is given by

$$\phi_{D_n}(t) = \int_t 2\pi f_{\mathrm{D}}(t)\mathrm{d}t = 2\pi f_{\mathrm{D}}t. \tag{2.30}$$

2.4.4 Delay Spread

Having introduced the pulse train, now we analyze the characteristic of the length of the pulse train, which has a different impact to signals with different bandwidths (W). First, let us introduce the metric that quantifies the length as *delay spread* (T_{S}). There are several characterization of the delay spread: mean delay spread ($\mu_{T_{\mathrm{S}}}$), RMS (root mean square) delay spread ($\sigma_{T_{\mathrm{S}}}$), maximum delay spread, which is taken as the time spread between the arrival of the first and last multipath signal.

Delay spread gives two distinct features of the signal: if delay spread T_{S} is $\ll W^{-1}$ then none of the rays are *resolvable*; two ray is resolvable only the time lag is greater than the inverse of bandwidth (W^{-1}). This is *narrowband fading* (aka flat fading), which means that the receiver sees the sum of the rays with different amplitude and phases as follows;

$$h(t) = \sum_{n=0}^{N(t)} \alpha_n(t)\mathrm{e}^{-j\phi_n(t)}. \tag{2.31}$$

This either enhances or diminishes the signal as seen in Fig. 2.21. If delay spread is larger than the inverse of the bandwidth, then it experiences *wideband fading* (aka frequency selective fading) and causes intersymbol interference (ISI). ISI is basically interference of one symbol to the successor symbols. Typically, it is mitigated with equalization, multi-carrier modulation, and spread spectrum. Note that still wideband fading may experience constructive and destructive effect due to unresolvable rays. Additionally, several parameters are introduced in addition to the delay spread, such as coherence bandwidth, Doppler spread, and coherence time in order to statistically characterize the wideband fading.

Statistical characterization of delay spread can be obtained from the autocorrelation function of the $h(\tau,t)$, defined as

$$R_h(\tau_1,\tau_2;t,t+\Delta t) = E[h^*(\tau_1,t)h(\tau_2,t+\Delta t)], \tag{2.32}$$

however, R_h can be simplified if the channel is WSSUS[9] as follows;

$$R_h(\tau,\Delta t) = E[h^*(\tau_1,t)h(\tau_2,t+\Delta t)], \tag{2.33}$$

[9] Wide Sense Stationary and Uncorrelated Scattering.

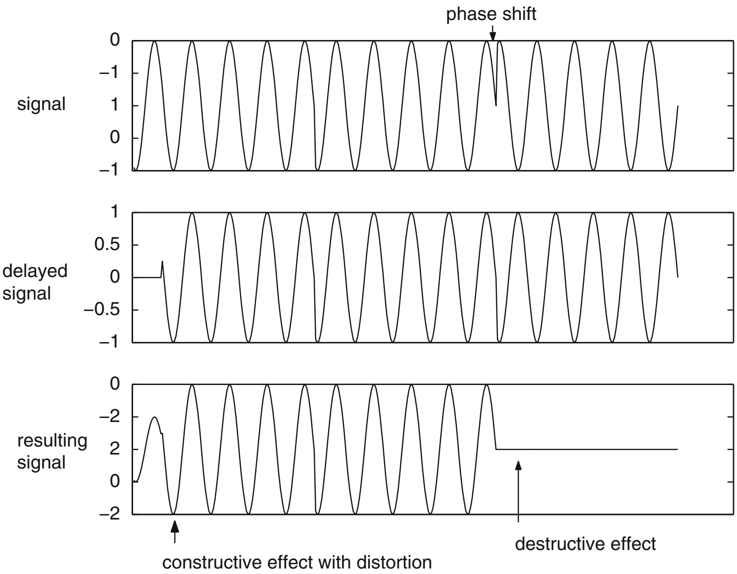


Fig. 2.21 Unresolvable rays

since WSS property states that joint statistics only depend on the time difference Δt not a particular time and US property states that two multipath component is uncorrelated at different delays $\tau_1 \neq \tau_2$. From this, mean (μ_{T_S}) and rms (σ_{T_S}) delay spread can be derived from $R_h(\tau, 0) \triangleq R_h(\tau)$ as follows

$$\mu_{T_S} = \frac{\sum \tau R_h(\tau)}{\sum R_h(\tau)} \tag{2.34}$$

and

$$\sigma_{T_S} = \sqrt{\frac{\sum(\tau - \mu_{T_s})^2 R_h(\tau)}{\sum R_h(\tau)}}. \tag{2.35}$$

In brief, delay spread gives a measure of the channel-time dispersion. We can further analyze the pulse train in the frequency domain to get a measure of the fading rate with changing frequency. This measure is termed coherence bandwidth.

2.4.5 Coherence Bandwidth

Coherence bandwidth (B_C) is a flatness measure of the channel in the frequency domain. Within that frequency interval, two frequency experience comparable or correlated amplitude fading. To analyze this we need the autocorrelation function of the $H(f,t)$, which is the Fourier transform of $h(\tau,t)$. $R_H(\Delta f,\Delta t)$ is the autocorrelation function of $H(f,t)$ since WSSUS property still holds. If we consider only the $R_H(\Delta f,0)$ then we can infer that B_C equals to Δf when $R_H(\Delta f,0) \approx 0$. This means that frequencies within a coherence bandwidth of one another tend to all fade in a similar or correlated fashion.[10] If we rephrase this, B_C gives minimum frequency separation that indicates the channel components are independent.

Note that Fourier transform of $R_h(\tau,0)$ is $R_H(\Delta f,0)$, which shows that delay spread and coherence bandwidth is related. For instance, in ideal communication, delay spread is zero and coherence bandwidth is infinite. Typically, B_C is where the correlation is $|R_{\mathrm{H}}(\Delta f,0) = 0.5|$ for two fading signal envelopes at two frequencies. This corresponds to $B_C \approx 1/(5\sigma_{T_{\mathrm{S}}})$. However, if correlation is 0.9 then $B_C \approx 1/(50\sigma_{T_{\mathrm{S}}})$. From this we can deduct that if the channel bandwidth (W) is greater than the B_C then some portion of the signal bandwidth experience frequency-selective fading due to ISI. This degrades performance considerably and system requires an *equalizer* or a way to avoid ISI such as multicarrier modulation or spread spectrum (We will be discussing in Chap. 4). However, if $W \ll B_C$ then fading is flat across the entire signal bandwidth as seen in Fig. 2.22. Note that flat fading does not require an equalizer.

Delay spread and coherence bandwidth are parameters that describe the *time dispersive nature*; however, they give no information about *time varying nature* of

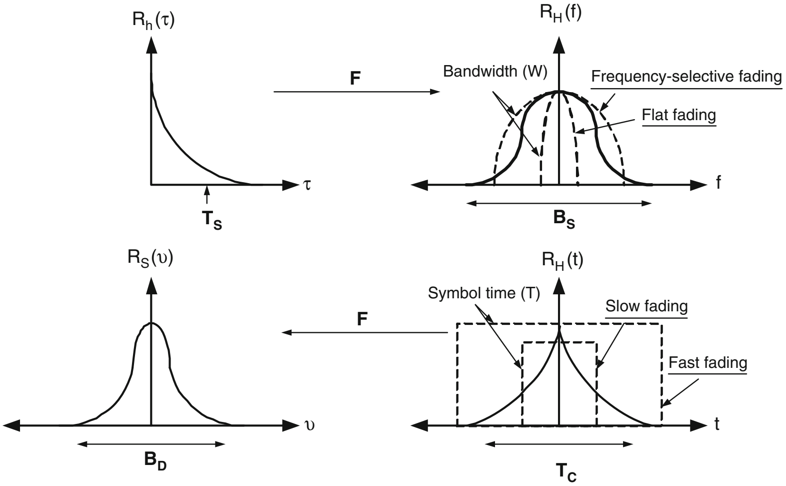


Fig. 2.22 Relationship between delay spread, coherence bandwidth, Doppler spread, and coherence time

[10] "One reason for designing the CDMA IS-95 waveform with a bandwidth of approximately 1.25 MHz is because in many urban signaling environments the coherence bandwidth B_{C} is significantly less than 1.25 MHz. Therefore, when fading occurs it occurs only over a relatively small fraction of the total CDMA signal bandwidth. The portion of the signal bandwidth over which fading does not occur typically contains enough signal power to sustain reliable communications...".

the channel due to mobility. Next, we introduce Doppler spread and coherence time as the parameters that describe the time varying nature of the channel.

2.4.6 Doppler Spread

We introduced the Doppler shift, which is the different rates of change in phase due to mobility. The difference in Doppler shifts that contributes to a single channel tap is known as *Doppler spread* (B_{D}), which is the measure of spectral broadening. In terrain mobile channels, Doppler spread of a narrowband signal is usually equal to the maximum Doppler shift i.e., the spectrum is spread over a band of f_{D}. In general, it can be derived from $R_{\mathrm{H}}(\Delta f, \Delta t)$ by taking the Fourier transform with respect to Δt:

$$R_{\mathrm{S}}(\Delta f, \upsilon) = \int_{-\infty}^{\infty} R_{\mathrm{H}}(\Delta f, \Delta t) \mathrm{e}^{-j2\pi\upsilon\Delta t} \mathrm{d}\Delta t. \tag{2.36}$$

To get the Doppler at the single frequency Δf is set to zero consequently maximum υ value that makes the $R_{\mathrm{S}}(\upsilon)$ ($\triangleq R_{\mathrm{S}}(0, \upsilon)$) greater than zero is the Doppler spread of the channel. If the baseband signal bandwidth is much greater than $W \gg B_{\mathrm{D}}$ then, the effects of Doppler spread are negligible at the receiver.

2.4.7 Coherence Time

From $R_{\mathrm{H}}(\Delta f, \Delta t)$, we can also deduct how channel decorrelates over time by setting Δf to zero. Range of Δt values that $R_{\mathrm{H}}(\Delta t)$ is zero is defined as channel *coherence time* (T_{C}), which is a measure for time varying nature the frequency dispersiveness of channel in time domain since it indicates the time duration, during which two signals have strong potential for amplitude correlation. It also implies that two signals arriving with time separation ($> T_{\mathrm{C}}$) are affected differently by the channel.

Coherence time is dual of Doppler spread in frequency domain since $R_{\mathrm{S}}(\upsilon)$ is Fourier transform of $R_{\mathrm{H}}(\Delta t)$. As a result, in general $B_{\mathrm{D}} \approx k/T_{\mathrm{C}}$ for some k. If coherence time is defined as the time over which correlation is above 0.5 then coherence time can be approximated as $T_{\mathrm{C}} = 0.423/B_{\mathrm{D}}$.

Previously, we classified the fading as flat vs. frequency selective, now we can also introduce slow vs. fast fading according to coherence time as illustrated in Fig. 2.22. *Slow fading* arises when the coherence time is larger than the symbol time (T) of the signal ($T \ll T_{\mathrm{C}}$). This makes the amplitude and phase of the channel almost constant over the period of use. Shadowing and rain fading are examples of slow fading. In *fast fading* channel, the coherence time of the channel is small relative to the symbol time ($T \gg T_{\mathrm{C}}$). As a result, the amplitude and phase of the channel change significantly over the period of use.

2.4.8 Channel Models

We introduced the narrowband fading model in (2.31). We can rewrite the equation as follows according to Clarke's formulation;

$$h(t) = h_I(t)\cos 2\pi f_c t - h_Q(t)\sin 2\pi f_c t, \tag{2.37}$$

where the in-phase and quadrature components are given by as follows;

$$h_{\mathrm{I}}(t) = \sum_{n=1}^{N(t)} \alpha_n(t)\cos\phi_n(t), \tag{2.38}$$

$$h_{\mathrm{Q}}(t) = \sum_{n=1}^{N(t)} \alpha_n(t)\sin\phi_n(t). \tag{2.39}$$

One can approximate $h_{\mathrm{I}}(t)$ and $h_{\mathrm{Q}}(t)$ as jointly Gaussian random processes by the central limit theorem when $N(t)$ is large. It is also a zero-mean Gaussian process since $E[r_{\mathrm{I}}(t)] = E[r_{\mathrm{Q}}(t)] = 0$. This is because $\alpha_n(t)$ and $\phi_n(t)$ are independent random process as a result $E[\alpha_n(t)] = 0$ and $E[\sin\phi_n(t)] = E[\cos\phi_n(t)] = 0$. We can show that $z(t) = |h(t)|$ is *Rayleigh* distributed since Z is Rayleigh random variable if it is composed of any two Gaussian variables X and Y with zero mean and variance σ^2 as follows; $Z = \sqrt{X^2 + Y^2}$.

Rayleigh distribution has the following probability density function

$$p_Z(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right), \quad r \geq 0, \tag{2.40}$$

where σ is RMS value of the received signal before envelope detection and mean value of Rayleigh distribution is given by 1.2533σ. An example for Rayleigh fading is illustrated in Figs. 2.23 and 2.24 shows the comparison between a Rayleigh fading channel and an AWGN channel.

Level crossing rate (LCR) and *average fade duration* of Rayleigh fading are two important metrics, which are useful for designing error control codes and diversity scheme used in mobile communication. LCR is related to the signal level and velocity of mobile and defined as the expected rate at which Rayleigh fading envelope, normalized to local RMS signal level, crosses the specified level in positive direction.

Note that this holds if $\phi_n(t)$ are uniformly distributed and this assumption is violated if there is a LOS component since $\phi_n(t)$ is dominated by the LOS component. Hence, $h(t)$ is not zero-mean. In this case, signal envelope shows a *Rician* distribution. Rician random variable is given by the following

$$p_Z(r) = \frac{r}{\sigma^2}\exp\left(\frac{-(r^2 + A^2)}{2\sigma^2}\right) I_0\left(\frac{Ar}{\sigma^2}\right), \quad A \geq 0, r \geq 0, \tag{2.41}$$

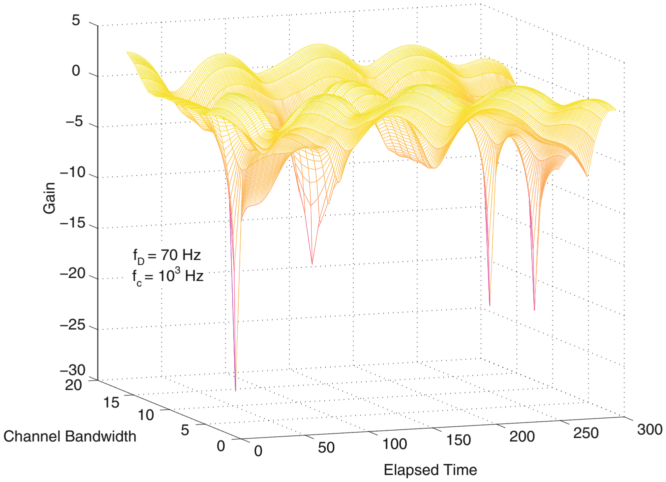


Fig. 2.23 Rayleigh fading

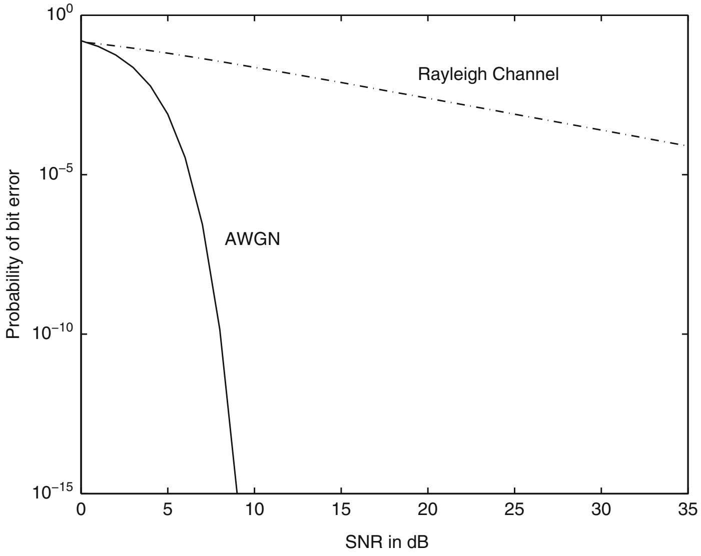


Fig. 2.24 Rayleigh fading on BPSK compared with AWGN ($E(\alpha^2) = 1$): Probability of error in Rayleigh fading is $P_e = \frac{1}{2}(1 - \sqrt{(\frac{\theta}{1+\theta})})$, where $\theta = SNR.E(\alpha^2)$ and α is Rayleigh distributed. P_e equals to $Q(\sqrt{(2SNR)})$ for AWGN channel

where A is peak amplitude of dominant signal and $I_0(.)$ is Bessel function of first kind. Rician distribution is described in terms of deterministic signal power and variance of multipath with parameter K as follows;

$$K = \frac{A^2}{2\sigma^2}, \tag{2.42}$$

where K basically states the power ratio of the LOS path over to the other non-LOS paths. If $K \rightarrow 0$ then the Rician RV converges to Rayleigh RV since dominant signal becomes weaker.

Rayleigh and Rician fading models consider infinite number of rays. However, there are empirical models that introduce fading models with finite number of multipath components for more real settings.

2.4.8.1 Stanford University Interim (SUI) Channel Models

SUI considers the terrain models introduced by the Erceg model. The multipath fading is introduced as a tapped delay line with 3 taps with nonuniform delays and Rician distribution gains with maximum Doppler frequency. There are six channel models for three terrain types (Terrain A, B, and C) as seen in Table 2.3.

Table 2.4 illustrates the SUI-1 channel model for a cell size of 7 km ($h_t = 30$ m, $h_m = 6$ m). Base station beam width is selected as 120° and two modes of beam width (omni directional 360° and 30°) are considered for the receive antenna. Note that channel gain need to be normalized before using the SUI model and Gain

Table 2.3 SUI channels

Channel	Terrain type	Doppler spread	Delay spread	LOS
SUI-1	C	Low	Low	High
SUI-2	C	Low	Low	High
SUI-3	B	Low	Low	Low
SUI-4	B	High	Moderate	Low
SUI-5	A	Low	High	Low
SUI-6	A	High	High	Low

Table 2.4 SUI-1 Channel Model for Terrain C

	Tap 1	Tap 2	Tap 3
Delay (µs)	0	0.4	0.9
Power (omni in dB)	0	−15	−20
90% K-factor (omni in dB)	4	0	0
75% K-factor (omni in dB)	20	0	0
Power (30° in dB)	0	−21	−32
90% K-factor (30° in dB)	16	0	0
75% K-factor (30° in dB)	72	0	0
Doppler (Hz)	0.4	0.3	0.5

Antenna Correlation: $\rho_{ENV} = 0.7$, Gain Reduction Factor: GRF = 0dB, Normalization Factor: $F_{omni} = -0.1771$dB, $F_{30^\circ} = -0.0371$dB, RMS delay spread: T_S(Omni/30°) = 0.111/0.042 µs, K (omni/30°): 3.3/14.0 (90%), 10.4/44.2 (75%)

Reduction Factor (GRF) is the total mean power reduction for a non-omni antenna compared with an omni antenna in dB, which should be added to the path loss.

2.4.8.2 ITU

ITU-R recommendation is also commonly used as an empirical channel model. ITU-R recommends six channels for three cases and two different delay spreads: indoor, pedestrian, vehicular with low delay spread (Channel A) and medium delay spread (Channel B). WiMAX Forum recommends Pedestrian A and Vehicular B channels as illustrated in Tables 2.5 and 2.6.

3GPP also considers ITU channel models. ITU Pedestrian A and Vehicular A channel models for LTE are used to represent the low, medium, and high delay spread environments with classical Doppler spread. Typical RMS values and maximum excess tap delay are presented in Table 2.7. The Doppler spectrum is modeled using Jake's Doppler spectrum. If $f_\mathrm{D} = f_c \frac{v}{c}$ denotes the maximum Doppler frequency and P is the net power, then power spectral density is given by

Table 2.5 ITU Channel model for pedestrian

Relative delay (ns)	Channel A (average power in dB)	Channel B (average power in dB)
0	0	0
110	−9.7	
190	−19.2	
200		−0.9
410	−22.8	
800		−4.9
1,200		−8.0
2,300		−7.8
3,700		−23.9

Table 2.6 ITU Channel model for vehicular

Relative delay (ns)	Channel A (average power in dB)	Channel B (average power in dB)
0	0	−2.5
300		0
310	−1.0	
710	−9.0	
1,090	−10.0	
1,730	−15.0	
2,510	−20.0	
8,900		−12.8
12,900		−10.0
17,100		−25.2
20,000		−16.0

Table 2.7 Delay profiles for LTE channel models

Tap delay/Extended	Pedestrian A EPA (45 ns RMS) (dB)	Vehicular A EVA (357 ns RMS) (dB)	Typical Urban ETU (991 ns RMS) (dB)
0	0.0	0.0	−1.0
30	−1.0	−1.5	
50			−1.0
70	−2.0		
90	−3.0		
110	−8.0		
120			−1.0
150		−1.4	
190	−17.2		
200			0.0
230			0.0
310		−3.6	
370		−0.6	
410	−20.8		
500			0.0
710		−9.1	
1,090		−7.0	
1,600			−3.0
1,730		−12.0	
2,300			−5.0
2,510		−16.9	
5,000			−7.0

$$S(f) = \begin{cases} \dfrac{P}{\pi f_{\mathrm{D}} \sqrt{1 - \left(\frac{f}{f_{\mathrm{D}}}\right)^2}} \end{cases} \tag{2.43}$$

for $|f| < f_{\mathrm{D}}$, otherwise $S(f)$ is 0. The LTE requirements state high, middle, and low Doppler frequencies for different mobility environments. Maximum Doppler frequency at $v = 350\,\mathrm{km/h}$ is $f_{\mathrm{D}} = 843\,\mathrm{Hz}$ for $f_c = 2{,}690\,\mathrm{MHz}$. On the other hand, f_{D} equals to 299 Hz for common high speed scenarios around 120 km/h with the same carrier frequency. f_{D} equals to 5 Hz for low mobile environments – speeds ranging from 2.3 to 7 km/h for any existing frequency band.

2.5 Diversity Techniques

We present the wireless channel as a challenge to overcome; however, diversity combining is a remarkable technique to leverage the independently fading signals to increase the capacity. The methodology relies on the wireless channel condition to create independent channels at least not to experience a deep fade in one of them. These realizations are combined in a way to get strongest signal.

- *Frequency diversity* carries the signals in different carrier frequencies far apart with each other. Frequency separation must be more than the coherence bandwidth to achieve uncorrelated signal fading.
- *Time diversity* sends the data over the channel at different times. Time separation is directly proportional to the reciprocal of the fading bandwidth, which is proportional to the speed of the mobile station.
- *Space diversity – Antenna Diversity* – uses multiple antennas in the receiver, which have distance in between to ensure independent fading. The separation around half-wavelength is ample to obtain uncorrelated signals.
- *Polarization diversity – Antenna Diversity* – utilizes the antennas either for a horizontal polarized wave or a vertical polarized wave. This can be a special case of space diversity and only two diversity branches can be possible.
- *Angle diversity – Antenna Diversity* – is achieved by directional antennas. The received signal arrives at the antenna via several paths, each with a different angle of arrival. The signals that are received from different directional antennas pointing at different angles are uncorrelated.

There are two type of diversity: microdiversity and macrodiversity. On the one hand microdiversity is combining within one BS to mitigate the multipath fading, macrodiversity, on the other hand, is combining the signals received by several base stations with coordination among them.

2.6 Multiple Access Schemes

So far, we introduced the communication mechanism from one source to a destination. Let us now consider the real situation where there are multiple sources that try to get *right to transmit* by using the spectrum resources. This is mediated with a multiple access scheme that facilitates the available resources. There are three basic dimensions to realize a resource: *frequency*, *time*, and *code*. Lately, *space* is also added with SDMA. We now introduce the basic multiple access schemes:

Frequency Division Multiple Access (FDMA) is based on splitting the frequency component of the spectrum into a number of channels where a user is allowed to transmit and receive in one of the channels. To prevent leakage from one channel into another, a frequency gap is introduced between each channel; however, this makes FDMA very inefficient. Of course, when a channel is free it is given to a new user. In spite of this, FDMA can only support few users concurrently and only used in *first generation* cellular systems.

Time Division Multiple Access (TDMA) splits time component of the spectrum into slots and a user is given a particular time slot, which may repeat periodically. This increases the capacity dramatically. GSM, the Global System for Mobile communications, which is a second generation cellular technology, employs both FDMA and TDMA. GSM splits the available spectrum into channels and assigns a channel to a cell. Within each cell, the frequency is used with TDMA principle.

Code Division Multiple Access (CDMA) system separates the users out by utilizing an orthogonal code per each user. Orthogonal codes are used to spread the energy of the signal over the frequency to allow multiple users to be separated at the receiver. The two most common forms are: frequency hopping (FH) and direct sequence (DS). FH is a switching technique over wide frequencies by a *hopping pattern* specified by the code. Hence, interference is minimized since at a particular instant in time, the interference is present in a narrow band only since overlaps are minimal due to orthogonality of the codes. FH scheme is used in IEEE 802.11b wireless LAN and Flash[11]-OFDM systems.

DS employs a binary modulation with a code, which is a pseudo-random sequence[12] of ±1. This spreads the signal and the same code is used to extract the original signal in the receiver. DS-CDMA is used in Wideband CDMA for the air interface of UMTS.

CDMA suffers from *near-far problem*, subscribers close to the base station may transmit with higher power that exceeds the average signal of others. Consequently, this causes hearability problem at the receiver for the users that are away from the base station. CDMA provides single frequency reuse, which is the ability to reuse the same channel at other cell sites. In the FDMA and TDMA systems, frequency planning is required to insure that the interference from adjacent cells is minimized. However, in CDMA, the channelization is performed using the codes by assigning different set of codes to adjacent cells to insure less correlation with the signal from a nearby cell. Also, note that CDMA system has the ability to perform *soft handover* (make-before-break), which is the mechanism that allows a mobile station to communicate simultaneously with two or more cells. This is different than *hard handover* (break -before -make) introduced in the beginning of the chapter.

Random Access scheme is another multiple access scheme, which is widely popular in WLAN and Ethernet because of its distributed fashion. Random access scheme grants channel with contention. Stations back off a random amount of time and then try to access the channel. If they detect a transmission, they back off again for another random time. Otherwise, they transmit. If there are more than one transmission, collusion occurs and stations perform the back off process again for retransmission. In WLAN, random access is built over TDMA since station uses the channel in time.

Wireless Token Ring Protocol (WTRP) is another distributed technique that reduces the unutilized channel when compared with random access. In WTRP, stations form a ring and a token packet determines right to transmit. A station who owns the token has right to transmit for a fixed/variable amount of time. Then, it passes the token to the successor in the ring. Again, it is built over TDMA.

[11] Flarion Low Latency And Seamless Handover.

[12] Pseudo-random (PN) sequences are random but deterministic sequences composed of equal number of $+1$s and -1s. Autocorrelation of PN sequence is at maximum value when perfectly aligned.

2.7 OFDMA

Orthogonal Division Multiple Access (OFDMA), which is the topic of this book, utilizes a form of FDMA, TDMA, and CDMA all together with the advantages of Orthogonal Frequency Division Modulation (OFDM). In a classical parallel data system, the total frequency band is divided into *N* nonoverlapping frequency subcarriers. Each subcarrier is modulated with a separate symbol and then the *N* subcarriers are frequency-division-multiplexed (FDM). To provide a better interchannel interference, spectral overlap is not recommended to avoid high-speed equalization and to combat impulsive noise and multipath distortion. However, this leads to inefficient use of the available spectrum. To cope with this inefficiency, the ideas proposed from mid-1960s were to use parallel data and FDM with overlapping subchannels as can be seen in Fig. 2.25: overlapping technique would give %50 more subcarriers but reduce the adjacent interference orthogonality between the different modulated carriers.

OFDM modulation creates a direct mathematical relationship between the frequencies of the carriers in the system so that the sidebands of individual carriers overlap and the signal is still received without adjacent carrier interference. Moreover, single high-rate bit stream is converted to low-rate *N* parallel bit streams where each can be modulated differently as seen in Fig. 2.26.

The receiver in OFDM system acts as a bank of demodulators, translating each subcarrier down to DC, with the resulting signal integrated over a symbol period (T). If the carrier spacing is a multiple of $1/T$ then other subcarriers in time domain have cycles when integrated result is zero. Thus, the carriers are linearly independent.

Orthogonal frequency carriers are created by Discrete Fourier Transform (DFT). Figure 2.27 shows the spectrum of the OFDM symbol and note that center frequency of each subcarrier shows no crosstalk from other channels. Using the DFT in the receiver and calculating the correlation values with the center frequency of each subcarrier can result in recover of the transmitted data. DFT-based technique achieves frequency division multiplexing by baseband processing not with bandpass filtering as common for FDMA and others.

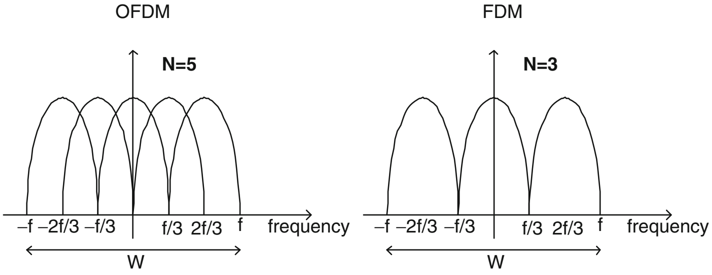


Fig. 2.25 OFDM vs FDM

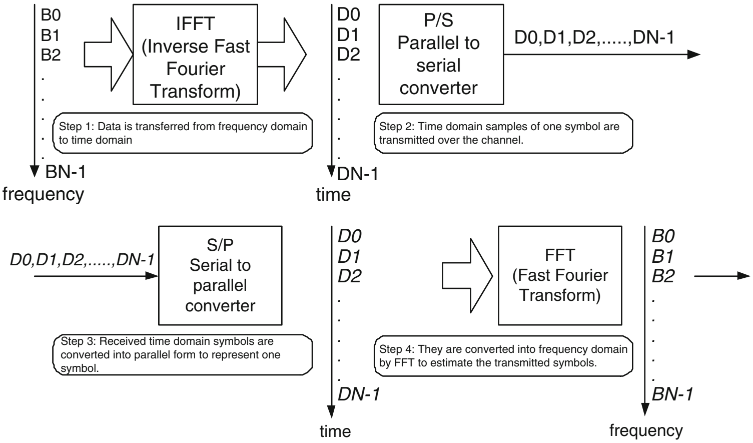

B0
B1
B2
BN-1
frequency
IFFT (Inverse Fast Fourier Transform)
Step 1: Data is transferred from frequency domain to time domain
D0
D1
D2
DN-1
time
P/S Parallel to serial converter
D0,D1,D2,.....,DN-1
Step 2: Time domain samples of one symbol are transmitted over the channel.
D0,D1,D2,.....,DN-1
S/P Serial to parallel converter
Step 3: Received time domain symbols are converted into parallel form to represent one symbol.
D0
D1
D2
DN-1
time
FFT (Fast Fourier Transform)
Step 4: They are converted into frequency domain by FFT to estimate the transmitted symbols.
B0
B1
B2
BN-1
frequency

Fig. 2.26 OFDM modulation concept

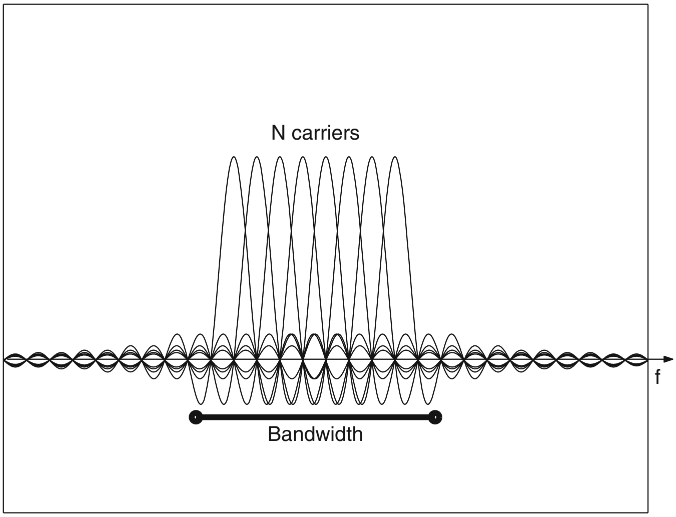


Fig. 2.27 Orthogonal carriers

Efficient implementation of DFT is fast Fourier Transform (FFT), which reduces the number of operations from N^2 in DFT to $N \log N$. Recent advances in VLSI[13] technology make high-speed, large size FFT chips commercially affordable. OFDM can be combined with multiple access using time, frequency, or coding separation of the users. OFDMA achieves multiple access by assigning different OFDM subcarriers to different users where assignment exists over a set of symbol time together with a hopping patterns that can be defined in a similar fashion as in CDMA to spread the interference.

[13] Very Large Scale Integration.

2.8 Duplexing: TDD, H/FDD Architectures

In mobile communication, another arrangement is needed to facilitate the communication between a mobile station and a base station. Note that this is a duplex communication and duplexing methods that separate uplink[14] (UL -mobile stations to base station) and downlink[15] (DL -base station to mobile stations) communication mainly fall into two main techniques: Time Division Duplexing (TDD) and Frequency Division Duplexing (FDD).

TDD utilizes single frequency channel for both transmission and reception. In effect, TDD divides the data stream into DL and UL frames. The radio design depicted in Fig. 2.28 illustrates that it requires only one Local Oscillator (LO) and RF filter is shared between transmitter and receiver. Note that having one synthesizer saves area on the silicon.[16] In TDD, transmission noise does not do self jamming to

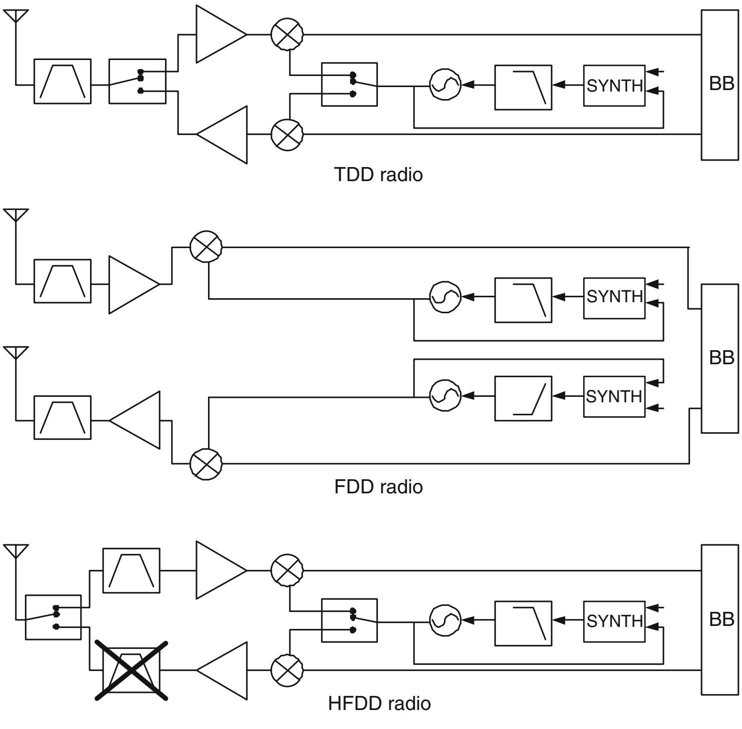


Fig. 2.28 Duplexing Radios Source: Intel Technology Journal, Volume 8, Issue 3, 2004

[14] aka reverse link (RL).

[15] aka forward link (FL).

[16] "Three areas of the radio are important for cost and power: synthesizer, power amplifier, and filter. The synthesizer generates the LO that mixes with the incoming RF to create a lower frequency signal, which is ready to be processed by the baseband. A high-performance synthesizer takes a die area and is therefore a costly component of the RFIC".

the reception since only one of them is on at any time. This also provides significant power savings but with some overhead since a guard band is required to separate the UL and DL, which increases overhead. TDD allows dynamic allocations for UL and DL, which is important for Internet-traffic since UL/DL ratio is no longer symmetric as in voice communication.

However, FDD radio requires two distinct frequency channel: one for DL and other one for UL together with a guard band in frequency that is required to separate the UL and DL. Consequently, in the radio design, it requires two LO and two RF filter as seen in Fig. 2.28. FDD incurs reduced latency when compared with TDD due to simultaneous UL and DL communication but note that UL/DL ratio is not dynamic anymore. Therefore, it has been preferred for voice-centric deployments where UL/DL ratio is 50/50.

To avoid the high design costs and power hungry system that FDD imposes on mobile stations, it can be desirable to use a hybrid duplex method called HFDD (Half-duplex FDD) in the subscriber and blend this with FDD base station. HFDD is very similar to TDD as seen in Fig. 2.28 and combines the benefit of TDD systems with frequency duplexing. An HFDD device transmits and receives at different times like a TDD device but uses different frequencies to communicate with an FDD base station. Therefore, HFDD station only uses half of the capacity of a full-duplex FDD subscriber.

The choice of TDD or FDD may be selected by the regulatory agency; each country and/or regulatory body can specify if one or more duplex methods are allowed for a licensed frequency band. Note that WiMAX has started with TDD mode and now been designing for H/FDD. LTE and UMB are introduced with H/FDD to comply to their 3G predecessors. Also note that TDD is the popular choice for unlicensed bands due to the inexpensive filters in the design.

2.9 Wireless Backhauling

Up to now, we talk about the means of communication between mobile stations and a base station. Now, we talk about how base stations are connected to the rest of the backbone network via backhauling. Backhauling is a general term that also define the connection of the DSLAMs in DSL to the nearest ATM or Ethernet aggregation node or the connection of the company sites in enterprise networking to metro Ethernet network. The backhauling technologies includes the following:

- DSL variants, such as ADSL, SHDSL, PDH, etc.
- SDH/SONET interfaces, such as E1/T1, E3, T3, STM-1/OC-3, etc.

and requires the following in general:

- High availability[17]
- Low latency

[17] requirements vary by operator from 99.9% to 99.999%.

- Scalability
- Network reach
- Path diversity

Alternative to traditional backhaul technologies like T1/E1, new way of backhauling is emerging via wireless to meet the increasing capacity demand for 4G networks.

- Point-to-point Microwave Radio Relay Transmission
- Point-to-multipoint Microwave Access Technologies, such as LMDS[18] WiMAX, WiFi, etc.

The important consideration in backhauling is *high availability* of the system. High availability is affected by the joint probability of two cojoined links failing simultaneously. For instance, service availability of 99.99% path becomes 99.995% if the link failure probability is 0.005%. Wireless backhauling addresses this high availability problem by introducing the flexibility of adding additional wireless links in various configuration. For example, Fig. 2.29 shows three different configuration options for wireless backhauling; *point-to-point*, *ring*, and *mesh*.

High availability in point-to-point link can be met with redundant equipment but still challenged by path availability. However, it is interrupted when path fails due to rain or other air link failures. Ring configuration is preferred over point-to-point link in some cases due to its 99.999% high availability since it meets this challenge

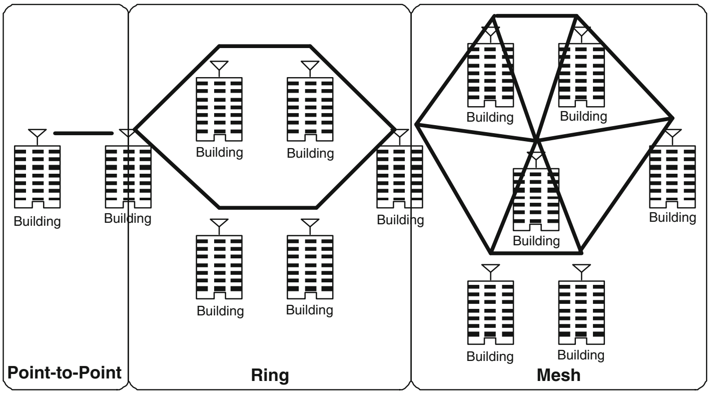


Fig. 2.29 Wireless backhauling options

[18] Local Multipoint Distribution Service (LMDS) is a broadband wireless access technology for last mile governed by IEEE 802.16.1 Task Group:

- LMDS operates on 26–29 GHz or 31–31.3 GHz microwave frequencies.
- Phase-shift keying or amplitude modulation is used.
- Links range up to 5 miles but limited to 1.5 miles due to fading in rain.
- LMDS frequencies can be utilized by point-to-point systems.

with path diversity while also inherently providing equipment redundancy. Mesh configuration reduces the link failure probability more by introducing alternate paths over ring configuration.

2.10 Summary

In this chapter, first we gave a brief overview of wireless cellular systems. Cellular system is introduced to combat with scarce spectrum by dividing the desired coverage area into smaller cells and reusing the same frequency in a spatially separated cells. This gives the way to introduce the desired coverage with limited spectrum; however, this structure requires handover, which is the transfer of communication from one cell to another without interruption. In addition to handover techniques, we introduced the following cell deployment strategies that are essential for the quality of session:

- Cell splitting
- Mico/macro cell sites
- Sectorization
- Femtocells and picocells
- Multi antenna systems

Second, we discussed the basics of digital communication in which we introduce the source and channel coding where former reduces redundancy in the information to transmit information with less bits and latter introduces redundancy to combat with error during communication. We also talked about error detection and error correction methods where former only detects the error and later also tries to recover the error. After that we discussed puncturing, interleaving and modulation techniques; puncturing adjusts the coding by removing some bits during transmission, interleaving changes the position of bits to distribute the error for better recovery. And, modulation packs the bits into waveforms. We also introduced hybrid automatic repeat request (HARQ), which is a hybrid acknowledgement scheme getting popular in today's communication systems.

Third, we discussed the fundamentals of the wireless channel. A transmission in the wireless channel may experience diffraction, scattering, or reflection that cause delayed replicas of the original transmission, which all superpose in the receiver with different amplitude and phase. Extended length of the transmission and shifts in the frequency due to mobility are important characteristic of the wireless channel. We introduced the following statistical parameters to characterize a wireless channel:

- Delay spread
- Coherence bandwidth
- Doppler spread
- Coherence time
- Flat and frequency selective fading
- Slow and fast fading

Finally, we introduced the evaluation of medium access techniques toward OFDMA and duplexing methods considered in cellular communication. At the end, we gave a brief introduction to newly emerging wireless backhauling.

References

1. ITU-R Recommendation M.1225, "Guidelines for evaluation of radio transmission technologies for IMT 2000," 1997.
2. Proakis, J. G., *Digital Communications*, McGraw-Hill, New York, 1995.
3. Goldsmith, A., *Wireless Communications*, Cambridge, 2005.
4. Winters, J.H., "Optimum Combining on Digital Mobile Radio with Co-channel Interference," *IEEE Sel Areas Commun*, no. 4, 1984.
5. Rappaport, T. S, *Wireless Communications*, Prentice Hall, New Jersey, 1996.
6. Kafle, P., *Digital Communications*, McGraW Hill, Inc., New York, 1995.
7. Walrand, J., Varaiya P., *High-Performance Communication Networks*, Morgan Kaufmann Publishers, San Francisco, 2000.
8. Levesque, A. H., Michelson, A. M., *Error Control Techniques for Digital Communication*, Wiley, New York, 1985.
9. Peterson, W. W., Weldon, E. J., Jr., *Error Correcting Codes*, 2nd ed. The MIT Press, Cambridge, MA, 1972.
10. Perez, L., Schlegel, C., *Trellis Coding*, IEEE Press, Piscataway, NJ, 1997
11. Wicker, S. B., *Error Control Systems for Digital Communication and Storage*, Prentice Hall, Englewood Cliffs, NJ, 1995.
12. McKown, J. W., Hamilton, R. L., "Ray tracing as a design tool for radio networks," *IEEE Network*, pp. 27-30, November 1991.
13. Steffen, A., "Secure Network Communication Part III," `http://www.strongsec.com/zhw/KSy_Auth.pdf`
14. Johnson, D., "Wireless Channels," `http://cnx.org/content/m0101/latest/`
15. Parsons, D., *The Mobile Radio Propagation Channel*, Wiley, New York, 1994.
16. Patzold, M., *Mobile Fading Channels*, Wiley, New York, 2002.
17. Bultitude, R. J. C., Bedal, G. K., "Propagation characteristics on microcellular urban mobile radio channels at 910MHz," *IEEE J Sel Areas Commun*, pp. 31–9, 1989.
18. Durgin, G., Rappaport, T. S., Xu, H., "Partition-based path loss analysis for in-home and residential areas at 5.85GHz," *IEEE Globecom*, pp. 904–9, 1998.
19. Jakes, W. C., *Microwave Mobile Communications*, Wiley, New York, 1974.
20. Haykin, S., *Communications Systems*, Wiley, New York, 2002.
21. Bello, P. A., "Characterization of randomly time-variant linear channels," *IEEE Trans.* vol. CS-11, no. 4, pp. 360–393, 1963.
22. Molnar, B. G., Frigyes, I., Bodnar, Z., Herczku, Z., "The WSSUS channel model: comments and a generalization," `docs4.mht.bme.hu/documents/wssus6/wssus6.html.`
23. Haykin, S., *An Introduction to Analog and Digital Communications*, Wiley, New York, 1989.
24. Simon, M. K., Alouini M.-S., *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*, Wiley, New York, 2002.
25. Andrews, J. G., Ghosh, A., Rias, M., *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*, Prentice Hall, 2007.
26. Jain, R., "Channel Models: A tutorial," `www.cs.wustl.edu/jain/cse574-08/ftp/channel_model_tutorial.pdf.`
27. Wilson, D. G., *Digital Modulation and Coding*, Prentice Hall, New Jersey, 1996.
28. Parsons, J. D., *The Mobile Radio Propagation Channel*, Pentech Press Publ., London, 1992.
29. Okumura, Y., Ohmori, E., Kawano, T., Fukua, K., "Field strength and its variability in UHF and VHF land-mobile radio service," *Rev Elec Commun Lab*, vol. 16, no. 9, pp. 825–73, 1968.

30. Hata, M., “Empirical formula for propagation loss in land mobile radio services,” *IEEE Trans Veh Technol*, vol. 29, pp. 317–325, 1980.
31. Erceg, V., et. al, “An empirically based path loss model for wireless channels in suburban environments,” *IEEE J Select Areas Commun*, vol. 17, no. 7, pp. 1205–1211, 1999.
32. Erceg, V., et. al., “A model for the multipath delay profile of fixed wireless channels,” *IEEE J Select Areas Commun*, vol. 17, no. 3, pp. 309–410, 1999.
33. European Cooperative in the Field of Science and Technical Research EURO-COST 231, “Urban transmission loss models for mobile radio in the 900- and 1,800MHz bands (Revision 2),” COST 231 TD(90)119 Rev. 2, The Hague, The Netherlands, 1991.
34. Ikegami, F., Yoshida, S., Takeuchi, T., Umehira, M., “Propagation factors controlling mean field strength on urban streets,” *IEEE Trans Anten Propag*, vol. 32, no. 8, pp. 822–829, 1984.
35. Walfisch, J., Bertoni, H. L., “A theoretical model of UHF propagation in urban environments,” *IEEE Trans on Anten Propag*, vol. 36, no. 12, pp. 1788–1796, 1988.
36. Boyle, K. R., “The performance of GSM 900 antennas in the presence of people and phantoms,” *Proc. 12th Int Conf Anten Propag (ICAP 2003)*, Exeter, UK, March 2003.
37. Char, A., Roberty, M., *Measurement Setup for 3G Phones*, Master’s thesis, Aalborg University, 2004.

Chapter 3
Basics of All-IP Networking

3.1 Introduction

Today's hierarchical architecture for cellular networks created in the circuit-switched era became inefficient in supporting real-time IP services. The shift to flat IP networks in cellular will deliver substantial cost and flexibility to mobile operation as well as address the increasing requirements of emerging applications.

This chapter is dedicated to examine the technology paths to the All-IP Network (AIPN) starting from basics of IP technology and continuing with advanced components of next-generation networks, which are widely used to structure WiMAX, LTE, and UMB standards.

Let us first look at IP in brief. How it is evolved? and what is the paradigm shift that is happening in mobile networking? In a communication network, devices are connected with a permanent physical link either wireless or wired and the physical link has a capacity that is shared by the users. Traditionally, in the circuit switching, a dedicated capacity is allocated to a voice connection and connection was kept alive even if no data are being transmitted. Typically, a connection is established with a delay but once connected, the link operates with a fixed data rate. Circuit switching is used in the public switched telephone network (PSTN) and private networks such as PBX[1] or private wide area network (WAN).

Packet switching evolved to handle digital traffic better than the circuit switching networks. Main idea is to share the capacity among users on *need* basis. The information is packetized and transmitted over the physical link as long as the link capacity permits. There is no end-to-end dedication but each packet has a header, which has the necessary information to route the packet to the destination. Intermediate entities buffer and forward the packets. This utilizes the link at maximum extent. Packet switching does not guarantee timely packet delivery

[1] "A private branch exchange (PBX) is a telephone exchange that serves a particular business or office, as opposed to one that a common carrier or telephone company operates for many businesses or for the general public...".

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_3, © Springer Science+Business Media LLC 2009

and introduces header overhead. Therefore, Internet has introduced advanced protocols to provide Quality of Service (QoS) for applications such as voice and video communication.

3.2 IP Protocol

Packet switching, which was started as a military project (ARPANET) to provide a robust communication architecture, has introduced routing methodology, which can immediately provide alternate paths when there is a node failure. Packet is routed according to the routing table of each router (an intermediate node).

Figure 3.1 shows that if "Router 2" fails, "Router 3" is alternate path for *Destination* B when *Source* is A. Alternatively, this structure creates more than one route to the destination. Router is responsible to buffer and forward the packet. Each router consequently has a capability for buffer storage and forward capacity. Routing methods aim to provide best-performance regarding capacity, latency, and robustness.

In IP protocol, a routing information *header* is appended to each data packet. Routing information header mainly contains the address of source and destination and available input for routing. IP Header includes *Type of Service* parameter to apply a priority when routing. This parameter helps to differentiate the service (DiffServ) and provide Quality of Service (QoS). IP header also contains source and destination addresses; an IP address consists of 32 bits for the current IP protocol (IPv4). By convention, it is expressed as four decimal numbers (32 bits) separated by

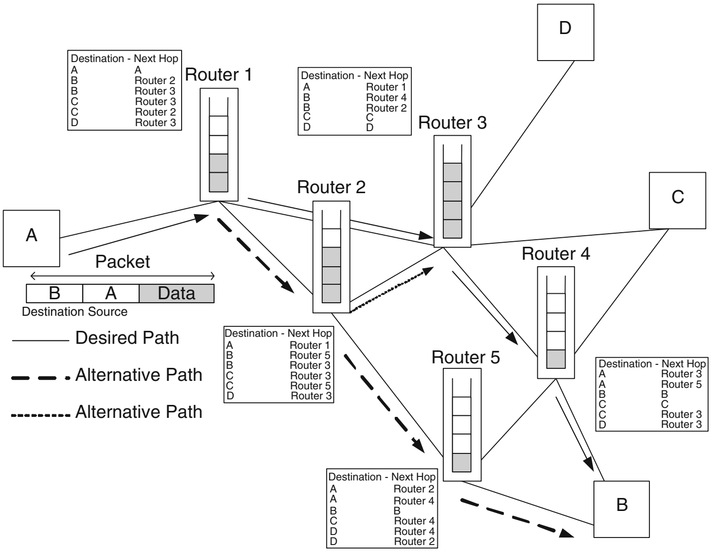


Fig. 3.1 IP network

Table 3.1 IP Address Classification

Class	Address range	Network address field size	Total network addresses	Host address field size	Total host addresses
A	1.0.0.0–126.0.0.0	7	126	24	16,777,214
B	128.0.0.1–191.255.0.0	14	16,383	16	65,534
C	192.0.1.0–223.255.255.0	21	2,097,151	8	254
D	224.0.0.0–239.255.255.255		28		
E	240.0.0.0–255.255.255.255				

full stops (*abc.def.ghl.klm*) and in theory, there can be 4,294,967,296 IP addresses. However, in reality, the number of IP addresses that can be assigned is less since taxonomy has been created within the IP address space to better handle the network with different sizes.

There are five classes as seen in Table 3.1. When trying to identify an address as class A, B, C, D, or E, the simplest way is to look at the address ranges since class A addresses range from 1.0.0.0 to 126.0.0.0, class B addresses start with 128.0.0.1 and ends with 191.255.0.0, class C addresses range from 192.0.1.0 to 223.255.255.0, class D address space, reserved for multicast, ranges from 224.0.0.0 to 239.255.255.255, and class E address space, reserved for special use, ranges from 240.0.0.0 to 255.255.255.255. For example "128.0.0.7" belongs to class B since "128." falls into class B range.

One can see from the table that each class can host different size of nodes. For example, a class A network can host 16,777,214 nodes but there can be at most 126 class A networks. However, a class B network can host 65,534 nodes but there can be 16,383 class B networks. This taxonomy, as can be seen from the table, is achieved by defining *network address field size* and *host address field size* in 32 bit IP address with different sizes.

Also, a single network may be subdivided into clusters via subnetting. Subnetting masks a portion of the *host address field* to introduce subnets and the remaining portion for the hosts (nodes). The general formulas to compute the number of possible subnets and hosts using a given subnet mask are given below:

$$\begin{aligned} \text{Subnet size} &= 2^{\#\text{masked}-\text{bits}} - 2 \\ \text{Host size} &= 2^{\#\text{unmasked}-\text{bits}} - 2\,, \end{aligned} \tag{3.1}$$

where $\#\text{masked} - \text{bits} + \#\text{unmasked} - \text{bits}$ equals to *host address field size*.

Typically, choice of IP address depends on network size; class A addressing is preferred for big networks; class B is suitable for average operators; class C is used for small networks. For example, a network for 2,000 host would use a class B address space. Note that this will leave 97% of address space unused since it can host 65,534 addresses. Also, if we use more than one class C networks to host that network, then the routing tables would be eight times bigger than before. To address this inefficiency, network address has been expressed with "IP address/mask size" tuple by CIDR (Classless Inter-Domain Routing) protocol. CIDR, introduced in 1993, is a new way of interpreting the IP addresses. It introduced variable-length subnet masking (VLSM) to introduce arbitrary-length prefixes.

For instance, 196.0.0.0/21 would route any address ranging from 196.0.0.0 to 196.0.7.0 to the same network since they have the same first 21 bits. This will provide a single entry instead of eight for 2032 host addresses. This also introduces flexible address block allocations to organizations based on their actual and short-term needs, rather than the very large or very small blocks required by *classful* addressing schemes introduced earlier. These issues have been addressed in IP version 6 (IPv6).

3.3 IP Address Assignment

Before introducing IPv6, let us first talk about IPv4 address assignment. We know that address space of IPv4 is limited and already a large set of IP addresses have been allocated. Typically, private address and unique ports are used along with Dynamic Host Configuration Protocol (DHCP) to assign an IPv4 address. DHCP introduces a method to lease an IP address for determined time period. If DHCP renewal is not initiated, that IP address is reused for another host.

The limited address space of IPv4 is further extended with private address space. Table 3.2 lists the private address space, which are only used within a network and not accessible from outside. Operator assigns a private IP address and a unique *port* as long as node resides in the ISP (Internet Service Provider) domain. To connect to Internet, there is a Network Address Translation (NAT) function, which has a public IP address, routable from outside. NAT is responsible to translate the private address/port to public address/port. As a result, many *private* IP addresses map to the *public* IP address and one-to-one mapping is maintained between ports. There can be 65,536 unique port number per IP address; consequently, a network size of 65,536 at maximum can be hosted with one public IP address.

Of course, NAT not only translates the private IP address into public IP address but also regenerates IP header checksums, TCP/UDP header checksums, and ICMP header checksum. Also, UDP and TCP port numbers and ICMP message types must be translated as well. Also binding between public and private IP addresses can be either static or dynamic. In dynamic NAT, binding is created on the fly and when the connection is terminated (or timeout), the address is returned to the pool for reuse.

To ease the use of Internet, rather than remembering the IP address for each destination, a distributed Domain Name Service (DNS) has been introduced to

Table 3.2 Reusable private addresses

Class	Addresses
A	10.0.0.0
B	172.16.0.0–172.31.0.0
C	192.168.0.0–192.168.255.0

translate the name (“http://www.berkeley.edu”) to IP address (“169.229.131.92”). DHCP assigns a DNS server to client and client establishes a DNS connection to translate the text entry to an IP address. The names are hierarchically administered in a decentralized manner via global coordination through ICANN.[2] For example, “**sem.etu.edu.tr**” is an address that points to a Web space belonging to an institute (**sem**) of TOBB Economy and Technology University (**etu.edu**) in Turkey (**.tr**). Authority to manage **.tr** domains is given to Turkey and TOBB ETU is also a local authority to append a left-most label to **etu.edu.tr** domain name.

3.4 IPv6

The first reason to introduce IPv6 was to address the insufficient number of available addresses in IPv4. On the one hand, IPv4 has a 4-byte addressing range giving a combination of $2^{32} = 4,294,967,296$ unique addresses. On the surface this seems like plenty of addresses, but the address classifications are very inefficient as explained in the previous sections. On the other hand, IPv6 supports $2^{128} \approx 3.4 \times 10^{38}$ addresses, which allows to allocate approximately 5×10^{28} IP addresses to each person living in the earth today. The key capabilities of IPv6, in comparison to IPv4, are as follows:

- IP address field is increased from 32 to 128 bits in length and incorporation of address hierarchy is supported,
- Simplified header format is introduced,
- Extension headers and options are supported,
- Authentication and privacy is provided,
- Auto-reconfiguration is provided,
- Incremental upgrade is possible,
- Low start-up costs is provided,
- Quality of service capabilities is provided,
- Mobility is supported.

The IPv6 header in Table 3.3 is 40 octets in length with eight fields (IPv4 has only 20 octets of header but 13 fields within it.). The type field identifies which IP protocol is in the payload (0×0800 for IPv4 and 0×86Dd for IPv6 [RFC1933]) to provide a smooth transition to IPv6. The simplest mechanism for IPv4 and IPv6 coexistence is to implement both protocol stacks on the same station. The station, which could be a host or a router, is referred to as an IPv6/IPv4 node, which has

[2] “The Internet Corporation for Assigned Names and Numbers (ICANN) is an internationally organized, nonprofit corporation that has responsibility for Internet Protocol (IP) address space allocation, protocol identifier assignment, generic (gTLD) and country code (ccTLD) Top-Level Domain name system management, and root server system management functions. These services were originally performed under US Government contract by the Internet Assigned Numbers Authority (IANA) and other entities. ICANN now performs the IANA function...”. Source: `http://www.icann.org`.

Table 3.3 IPv6 header

Header	Comment
Version	4-bit IP version number (=6)
Traffic class	8-bit traffic class field
Flow label	20-bit flow label
Payload length	16-bit integer to indicate length of payload in bytes
Next header	8-bit selector. Identifies the type of header immediately following the IPv6 header
Hop limit	8-bit unsigned integer. Decremented by 1 by each node that forwards the packet
Source address	128-bit address of the originator of the packet
Destination address	128-bit address of the intended recipient of the packet (possibly not the ultimate recipient, if a routing header is present.)

the capability to send and receive both IPv4 and IPv6 packets and can therefore communicate with an IPv4 station using IPv4 packets and an IPv6 station using IPv6 packets.

IPv6 hosts can be configured with Stateless Auto AutoConfiguration (SLAAC) or stateful configuration (DHCPv6) or manual configuration. In SLAAC, a host connects to a IPv6 network using ICMPv6 router discovery messages. Then, SLAAC lets a host send a link-local multicast solicitation request for its configuration parameters. Routers respond with a router advertisement packet that contains configuration parameters.

3.5 IP Transmission

So far we introduced the IP addressing scheme, now we briefly talk about IP transmission. To initiate an IP session, a client creates a connection of two types: connection-oriented (TCP) or connectionless (UDP). These connections are negotiated with peers not intermediate nodes and specified by *ports*.

TCP (Transmission Control Protocol) provides reliable and in-order delivery of packets with automatic repeat request (ARQ), which acknowledges the packets when received. If not received, server resends the same packet. Of course, retransmission introduces delay but overall it is reliable. TCP is used by WWW, E-mail, FTP, SSH, and some streaming media applications. However, because TCP is not optimized for timely delivery but accurate delivery, TCP is sometimes inefficient for real-time applications such as VoIP. For such applications, protocols like the Real-time Transport Protocol (RTP) running over the User Datagram Protocol (UDP) are usually recommended instead.

UDP is developed for connections where sporadic loss of data is bearable but untimely reception is not. As a result, there is no ARQ; consequently, there is no retransmission. However, UDP also alone does not guarantee timely delivery as well. Typically, RTP over UDP is used with the Real-time Transport Control Protocol (RTCP) for voice and video applications; RTP is used to transmit data and RTCP is used to monitor QoS required by voice and video applications.

After creating a connection, an IP packet is created with its IP header to be transmitted to the next hop specified in the routing table of the IP layer. The transmission to the next hop is typically over *Ethernet* protocol. Ethernet uses Address Resolution Protocol (ARP[3]) to find out the network interface that is linked to the next hop.

Network interface is identified by a globally unique 48-bit MAC address (hardware address). Client ARP transmits *ARP request* to find out the MAC address of the node. *ARP response* sent by next hop or any other node in the network contains the MAC address for the desired node. Then the client forwards the packet to the medium with an appended MAC header that has source and destination MAC addresses where source MAC address is the MAC address of the client and destination MAC address is the MAC address of the next hop received from the *ARP response*. This process is repeated in each intermediate node until the destination node. Selection of intermediate nodes (routers) are selected by the IP routing protocol.

3.6 IP Routing Protocols

IP routing protocols define the way the packets are forwarded and routed to the destination. Routers forward the packet to another router closer to the destination, according to their routing table. Routing table entries are created statically for small networks or dynamically for big networks. Dynamic routing table update requires information exchange between entities.

Routing protocols are classified into two groups as seen in Fig. 3.2: internal and external. Internal Gateway Protocols (IGP) such as Routing Information Protocol (RIP) and Open Shortest Path First (OSPF) carry out routing within a network managed by a single operator. External Gateway Protocols such as Border Gateway Protocol (BGP) handle routing between autonomous systems and are used, for example, to connect the gateways belonging to different ISPs on the Internet.

Interior Gateway Protocols are based on two algorithms: Bellman-Ford Algorithm or Dijkstra's Algorithm.

Bellman-Ford Algorithm [RFC1058] finds the shortest path to a destination. The nodes broadcast the routing tables every 30 s. The algorithm assumes that each node knows the length of the links attached to itself. Every node keeps track of current estimate of length to the destination. It updates if the new estimate is strictly

[3] "ARP is used in four cases:

- Two hosts are on the same network
- Two hosts are on different networks and must use a gateway/router to reach the other host
- A router needs to forward a packet for one host through another router
- A router needs to forward a packet from one host to the destination host on the same network"

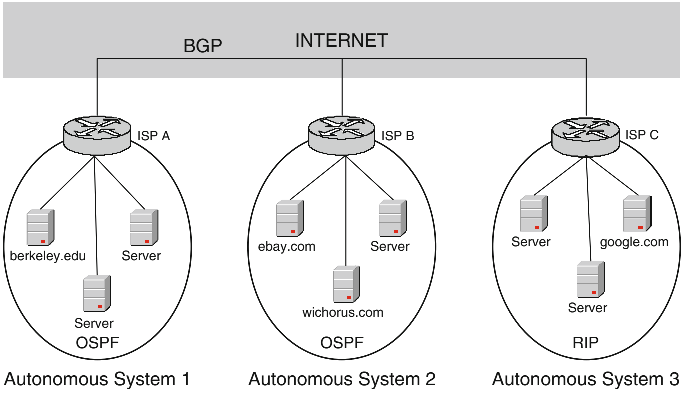


Fig. 3.2 Routing protocols

smaller than the previous estimate. Initially each node sets its estimate to infinity. Destination first sends message to its neighbors, which then update their estimates. Eventually, every node finds out the shortest distance to the destination, which does not need to be unique. The algorithm is also called *Distance Vector Routing* since each node sends to its neighbors the estimate of its distance to all nodes. It is a distributed algorithm and may fail to converge. The most common implementation is Routing Information Protocol (RIP). Interior Gateway Routing Protocol (IGRP) and enhanced IGRP are also common examples as well.

Dijkstra' Algorithm is a centralized algorithm where each node sends to all other nodes a message that indicates the length of the link. For this reason, it is considered a *Link State Routing* protocol.

Each node label is set to infinity but *root* node is labeled "0." Each node compares the label of each neighbor to the sum of the node's label plus the link's length. If that sum is smaller, it becomes the new label and the link is added to the lists of preferred links while the previous one is removed. Conceptually, the algorithm is called Open Shortest Path First (OSPF) protocol. Intermediate System to Intermediate System (IS-IS) protocol is also another link-state routing protocol designed by DEC.

3.6.1 RIP Version 2

The RIP1 (version 1) protocol [RFC1508, RFC1388] has a way to confine the network size by limiting the number of hops between source and destination to 15. Every node in the network advertises itself every 30 s. This requires frequent talk in the network. RIP also suffers from slow convergence since if there is a delay in the network it is high likely to achieve convergence more than 7 min (30 s × 15 hops).

The RIP2 [RFC2453] addresses the shortcomings of RIP1. RIP1 requires all subnets in a network class to be of the same size. RIP2 has variable length subnetting, which supports CIDR [RFC2082]. For backward compatibility RIP2 also supports 15 hops and does not provide authentication.

The RIP next generation (RIPng) [RFC2080 and RFC2081] is introduced for IPv6. RIPng also uses Bellman-Ford distance-vector algorithm to determine the best route to a destination. It is a distinct routing protocol from RIPv6 with no backward compatibility.

3.6.2 OSPF

OSPF [RFC1245, RFC1246, RFC1247] based on link state protocol is the most common IGP. When a new router is attached, it first discovers neighbors by sending *hello* packets to learn their network addresses. Then, it measures the delay or cost to reach each neighbor and sends a message, which contains this delay and cost information to all routers not just the neighbors as in RIP. Last, it computes the shortest path to every other router.

When compared with RIP, OSPF network is not confined to 15 hops. OSPF network can detect changes easily and has faster convergence time. OSPF reacts if there is a change in routing tables or every 30 min unlike RIP, which has frequent updates of 30 s. OSPF supports variable-length subnet masks and uses IP not TCP or UDP. However, RIP is simple and less complicated compared with OSPF.

OSPF development began in the very late 1980s, when memory was expensive, router performance was low and latency was high. However, with OSPFv2 restrictions based on early networking realities has been accommodated. Also, a new and improved version of OSPF (OSPFv3) [RFC2740] has been introduced to support IPv6. Similar to the relationship of RIPng to RIPv2 that OSPFv3 is not backward compatible with OSPFv2.

3.6.3 BGP Version 4

Border Gateway Protocol (BGP) [RFC4271] is an exterior gateway protocol (EGP), designed to exchange route information between different Autonomous Systems (AS), which is a network or group of networks under a common administration. Each AS is identified by its own AS number and its IP addresses.[4]

[4] Given by IANA and three Regional Internet Registries (RIRs):

- ARIN: American Registry for Internet Numbers is serving the Americas and part of Africa
- LACNIC: Latin American and Caribbean Internet Addresses is serving Latin America and Caribbean region
- RIPE NCC: Rseaux IP Europens Network Coordination Centre is serving Europe, the Middle East, the former Soviet Union, and the rest of Africa
- APNIC: Asia Pacific Network Information Centre is serving the Asia-Pacific region

In general, a BGP system exchanges reachability information with other BGP systems. This information includes information of the list of ASs that reachability information traverses and is used to construct a graph of AS connectivity. This way loops can be eliminated and some policy decisions at the AS level may be enforced. BGP supports CIDR and is based on Path Vector Protocol, which is a similar but complicated version of Distance Vector Protocol. BGP can be deployed within AS as well and called Interior BGP otherwise it is called External BGP.

BGP is suitable to provide *multihoming* in which single IP address is used for multiple links. Multihoming requires a public IP address range and an AS number. Then, a connection to two or more ISPs is established with control of a BGP enabled router. If an outgoing link fails, outgoing traffic will automatically be routed via one of the remaining links. Also, through BGP, other networks will be notified to reroute incoming traffic toward active links.

3.6.4 Multicast IP

Multicast IP [RFC1112, RFC1584] is a way of distributing the IP packet to a numerous hosts, which is identified by a group address. Internet Group Management Protocol (IGMP) implements a polling process for multicast routers to understand which group member is active. Multicast router, a designated IP router, sends a periodic *query* message to its hosts to ask which multicast group they belong. Hosts respond to the *query* messages by sending the *IGMP report* messages to indicate their group membership. All routers receive the *report* messages to note the memberships of hosts on the link.

3.7 QoS for All-IP Network

All-IP Network architecture requires implementing an end-to-end QoS, which shall consider three interfaces: air, access network, and core network. Air interface QoS is defined by the standards and executed by the base station and access gateway. QoS for IP-based access and core network on the other hand is defined with protocols as defined later.

3.7.1 DiffServ: Differentiated Services

DiffServ [RFC2474] provides a simple and coarse method to classify services of various applications. It creates traffic classes and rules to treat each class. Differentiated Services Code Points (DSCP) are defined in the IP header by replacing the TOS (Type of Service) byte as in Fig. 3.3. DSCP specifies the per-hop behavior (PHB),

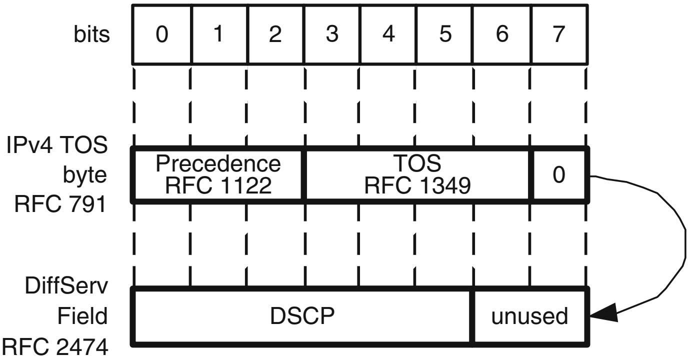


Fig. 3.3 Differentiated Services Code Points (DSCP) replaces the IPv4 Type-Of-Service byte. DSCP contains class selector codepoints and Per-Hop-Behaviors, which preserve IP precedence bits but TOS

where default PHB is set to (000000) for best effort traffic. There are currently two standard Per-Hop-Behaviors (PHBs);

- **Expedited Forwarding** (EF) [RFC2598] aims to minimize delay and jitter. Packets are treated with highest priority. It has a single DS codepoint (101110) and excess EF traffic is dropped.
- **Assured Forwarding** (AF) [RFC2597] has 12 DS codepoints with four classes and three drop-precedences (low/medium/high) within each class. Excess AF traffic may be demoted but not necessarily dropped.

PHBs are applied at a network ingress point (network border entry) or at originator and unmarked at the network egress point (network border exit). In between, they are routed according to the marking. DiffServ assumes Service Level Agreement (SLA) between networks. Out-of-profile traffic at the ingress has no QoS guarantee since policing and smoothing according to SLA at the egress point is expected. DiffServ provides simple and flexible resource utilization.

3.7.2 IntServ: Integrated Services

IntServ [RFC1633] is a signaling protocol to inform routers about the required QoS level for the session. IntServ provides finer-grained QoS compared to DiffServ. It was motivated by remote video, visualization, virtual reality, and multimedia conferencing. IntServ uses Resource Reservation Protocol (RSVP) to explicitly signal the QoS needs of an application. The idea is every application makes individual reservation and *FlowSpec* is introduced to define what the reservation is. FlowSpec consists of two parts: Traffic Specification (TSPEC) and Request Specification (RSPEC).

TSPEC defines the arriving rate and depth of tokens based on the token bucket algorithm. A token bucket fills up with tokens, arriving at constant rate. The arriving rate of the tokens dictate the rate of the flow and depth of the bucket dictates the burstiness off the flow.

RSPEC defines the requirements of the flow; *best effort* does not require reservation; *controlled load* guarantees that high percentage of packets are delivered, and transit delay suffered by packet is not significantly higher than the minimum transit delay, *guaranteed load* gives absolutely bounded service where the delay and number of packets dropped never goes beyond the specified limits.

IntServ requires several functions:

- End-to-end signaling
- Admission control
- Classification
- Policing
- Scheduling

IntServ is suitable for voice and data integration in small areas since it requires many states to be stored. DiffServ over RSVP provides scalable solution for networks with higher bandwidth. IntServ and DiffServ can exist in multi-tier approach where resource reservation per flow is performed in the edge network and in the core network, resources are reserved for aggregate flows only.

3.7.3 RSVP: Resource Reservation Protocol

RSVP [RFC2205] is a transport level control protocol to reserve certain QoS along the path for different set of applications. RSVP is not a routing protocol but works along with a routing protocol. A certain application residing in a host can initiate RSVP to request an appropriate level of service from network as seen in Fig. 3.4. Applications may differ in terms of required QoS. Certain applications require reliable delivery but timeliness of delivery is not important and other applications may require the opposite. For example, applications may be classified as follows:

- *Best Effort* traffic requires reliability but not timeliness of delivery. Applications include file transfer, disk mounts, interactive login, etc.

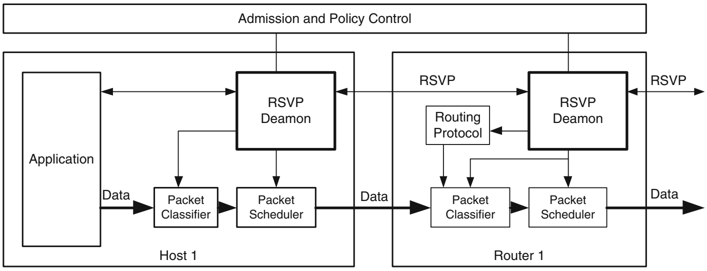


Fig. 3.4 RSVP protocol

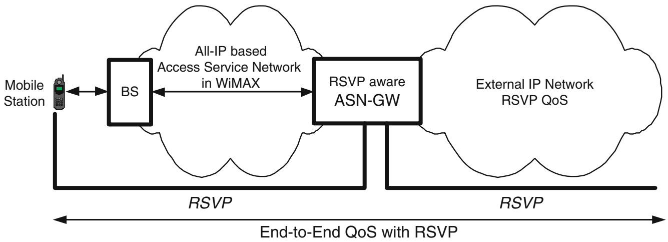


Fig. 3.5 End-to-end QoS in WiMAX with RSVP: Base stations in WiMAX are connected to ASN-GW, which deploys control and data plane functionality. ASN-GW is responsible to manage the QoS in the Access Service Network of WiMAX (More details is given in Chap. 10) and facilitate an interface to the external IP network

- *Rate Sensitive* traffic requires end-to-end guaranteed transmission rate. H.323 videoconferencing[5] is one application that requires constant bit rate.
- *Delay Sensitive* traffic requires timeliness of delivery with variable bit rate. MPEG-2 video is one of these applications since it is delta modulated at each frame so transmission rate changes; however, timely frame delivery is required.

An overview of RSVP operation is illustrated in Fig. 3.4. When request comes from application, *RSVP deamon* consults admission and control policy to ensure the availability of resources and required authority to make reservations. Required QoS is determined in the packet classifier and it is fed into the packet scheduler for transmission in orderly fashion regarding the QoS requirement. RSVP deamon communicates with the routers along the way to reserve the resources for the session with *RSVP path message*. A RSVP-aware router, in the mid-way, forwards the *RSVP path message* to the next hop toward the destination if it grants resources for this session. Destination replies with *Reservation-request message*, which is relayed back to the host. RSVP is simpler to establish paths for unicast and multicast traffic and requires two separate RSVP sessions for upstream and downstream.

RSVP can play a role in WiMAX between User and Access Service Network (ASN) Gateway (GW) as seen in Fig. 3.5. User can establish a Resource Reservation via ASN-GW, which needs to be RSVP aware to reserve bandwidth at the external network. RSVP context can be tunnelled in the ASN and only user and ASN-GW are aware of the context. If ASN-GW is not RSVP-aware then policy enforcement and resource reservation can not happen.

3.7.4 MPLS: Multiprotocol Label Switching

IP-based networks typically lack the quality-of-service features available in circuit-based networks, such as Frame Relay and ATM. MPLS[6] brings the sophistication

[5] H.320 in ISDN, H.310 in ATM.

[6] "MPLS stands for Multiprotocol Label Switching; multiprotocol because its techniques are applicable to ANY network layer protocol, of which IP is the most popular...".

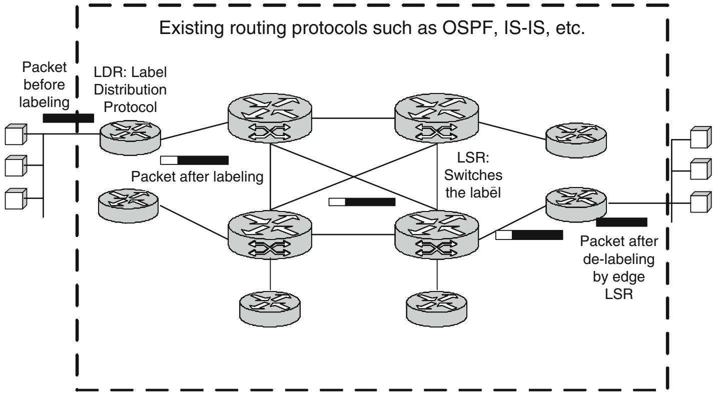

Existing routing protocols such as OSPF, IS-IS, etc.
Packet before labeling
LDR: Label Distribution Protocol
Packet after labeling
LSR: Switches the label
Packet after de-labeling by edge LSR

Fig. 3.6 MPLS Operation: MPLS has 32-bit header which contains the label (20-bits), the Class of Service (CoS) field (3-bits) to implement service classes, the Stack (S) field (1-bit) to support hierarchical label stack for routing packets through LSP Tunnels, TTL (time-to-live) field (8-bits) as in conventional IP TTL

of a connection-oriented protocol to the connectionless IP world. On the basis of simple improvements in basic IP routing, MPLS brings performance enhancements and service creation capabilities to the network.

There are Label Edge Routers (LERs), which labels each packet when they enter an MPLS network as in Fig. 3.6. These labels include destination, bandwidth delay, other QoS metrics as well as source IP address, socket information and differentiated service information. Once labeled, they are assigned to the corresponding Labeled Switch Paths (LSPs), where Label Switch Routers (LSRs) do label swapping.

MPLS defines Forwarding Equivalence Class (FEC) where IP packets belonging to the same class are forwarded over the same path with the same treatment. The assignment of a particular packet to a particular FEC is done once and FEC is encoded as a label. When the packet is forwarded, the label is sent along with it. In MPLS network, there is no further analysis of packet's network layer header since it is routed with the label only. The label is updated in the node and the packet is forwarded to next hop.

MPLS brings the Layer 2 switching speed to Layer 3 and routers make decision based on the content of the label rather than doing complex route lookup.[7]

Furthermore MPLS introduces the following:

- Provides classification of packets based on the network layer information to assign a priority as in Frame Relay and ATM type quality-of-service with CoS field
- Allows different level of traffic encryption since packet payloads are not examined
- Improves traffic engineering since a route is not necessarily automated although explicit route assignment is possible to force QoS

[7] "This was initial justification for MPLS but state-of-art Layer 3 switches perform route lookups at sufficient speeds with ASIC technology."

- Integrates the networks that operates either with Layer 2 or Layer 3 types of connectivity since MPLS is applicable to any network layer

DiffServ and MPLS are two different ways of addressing the IP QoS. DiffServ utilizes IP Type-Of-Service (TOS) field and replace DS byte to carry information about QoS level. Since it is a Layer 3 protocol, MPLS, on the other hand, maps the Layer 3 traffic to connection-oriented Layer 2 protocols by adding label for specific routing information. The main difference is as follows; DiffServ relies on edge entities in network (e.g., ASN-GW in WiMAX) to mark the packets. On the other hand, MPLS requires label-switching routers in the network.

3.7.5 DPI: Deep Packet Inspection

Another form of QoS which is becoming an essential part of the next generation network is DPI technology. DPI helps to prevent certain application traffic from unduly capturing resources and contributing to congestion. DPI (aka Intrusion Prevention Service) is also becoming the indispensable security feature to detect Viruses, Worms, Trojans, Spyware, Phishing before entering the network and enables ISPs to comply with CALEA[8] reporting requirements.

The standard way of packet inspection extracts basic protocol information revealed in the IP header. However, application-related deductions are impossible with basic inspection process. DPI inspection delves into the content of a TCP/UDP flow for a complete view. This requires reassembling of IP datagrams, TCP datastreams and UDP packets on runtime. For instance, certain flows can be rate limited according to predefined policies of the network. This ability will enable ISPs to offer new, differentiated services to subscribers with guaranteed QoS.

Typically, flows are identified according to the signatures. Signatures can be considered as the fingerprint of an application or protocol. Although a signature is developed to uniquely and completely identify the application or protocol, most of the cases the signature is not robust. This is because certain applications especially peer-to-peer (P2P) ones change pattern and apply encryption to remain unidentified. This makes the DPI problem challenging.

There are several methods identified for signature construction. *Port analysis* classifies the applications according to the ports they use. For instance, SMTP uses port 25 and HTTP uses port 80. However, some applications may select random ports or hide themselves as HTTP traffic. *String match* is another way, which looks at certain strings within the packets. *Numerical properties* look at the packet size and number of packets in the transaction or interarrival time of packets. These are all simple classification methods and may lead to many false positives since many P2P applications (eMule, BitTorrent, Skype) use encryption for obfuscation purposes to prevent content based inspection.

[8] "The Communications Assistance for Law Enforcement Act (CALEA) is a United States wiretapping law passed in 1994 (Pub. L. No. 103-414, 108 Stat. 4279, codified at 47 USC 1001-1010)."

Behavioral and *heuristics* analysis are getting the ultimate way of providing DPI functionality especially for P2P applications. Behavioral analysis is the classification according to the way a protocol acts and operates and heuristic analysis classifies based on statistical parameters. For example, packet size distribution is a behavioral pattern and average packet size is a heuristic pattern. Average number of ports used or entropy of the content is a heuristic pattern on the other hand switching from UDP to TCP ports on the run is a behavioral pattern.

DPI is becoming the essential part of cable, DSL, satellite, WiFi, WiMAX, and other networks. It is expected to introduce new sources of revenue by setting up per subscriber SLAs/policies and utilize the network more efficiently.

3.8 IP Header Compression

Typically, an IP session can be utilized more efficiently with header compression that removes the header field redundancies in packets belonging to the same flow to transmit only a few bytes of header information per packet since most of the header portion of the packet remains same during a data session. For instance, IP addresses of the client and the server remain the same and other fields such as sequence numbers change predictably. In many applications like VoIP and gaming, the header is almost of the same size or bigger than the payload so header compression is crucial.

IP Header compression provides solution to compress those headers over just one link, compressed at one end, uncompressed at the other end of the link. Savings can rise to 90% for some packets in addition to reduction in packet loss and increased turn around time.

The notion of header compression was introduced by Van Jacobson [RFC1144] in 1990 for the purpose of improving interactive terminal response of low-speed serial modem links. The original **Van Jacobson compression** scheme uses delta compression to compress the TCP header. Delta modulation sends only the difference of the changing fields. On the average, compressing 40 bytes to 4 is possible for TCP packets through CTCP (Compressed TCP Header). This requires no signaling between sender and receiver and relies on the ARQ of TCP for error recovery or residual bit errors. In the wireless setting, TCP ARQ introduces significant delays.

In 1999, Degermark, Nordgren, and Pink developed IPHC (IP header Compression) [RFC2507] techniques for compressing permutations of IPv4 headers, IPv6 base and extension headers, TCP and UDP headers, and encapsulated IPv4 and IPv6 headers for transmission over lossy links. That same year Casner and Jacobson added a specification for compressing IP/UDP/RTP headers commonly referred to as CRTP (Compressed RTP header). These are based on a mechanism similar to delta compression and introduced their own feedback mechanism to reduce error in transmission. Table 3.4 shows compression for VoIP packets. Note that there is almost 50% reduction in the packet size.

Table 3.4 VoIP packet with header compression: HC and GMH stand for header compression and generic MAC header, respectively

	G.729 with HC	G.729 without HC	AMR with HC	AMR without HC
Voice payload in bytes (inactive/active)	0/20	0/20	7/33	7/33
Headers in bytes (IPv4/IPv6)	2/4	40/60	2/4	40/60
>RIP		12		12
>UDP		8		8
>IPv4/IPv6		20/40		20/40
802.16e GMH	6	6	6	6
CRC	4	4	4	4
Packet size when inactive (IPv4/IPv6)	0/0	0/0	19/21	57/77
Packet size when active (IPv4/IPv6)	32/34	70/90	45/47	83/103

ROCCO[9] (Robust Checksum-based Header Compression) is proposed to improve performance over lossy links with long round-trip times (RTT) by a group of researchers from Ericsson and Lulea University as an Internet Draft in June 1999. ROCCO was designed to perform well with audio and video codecs, which tolerate bit errors in data frames. Studies showed that ROCCO compressed headers to half of the size of those provided by CRTP and remained robust over links with BER rates an order of magnitude higher.

In July 2001, the Robust Header Compression (ROHC)[10] was presented as a new standard for header compression by a working group of the IETF [RFC3095]. The Robust Header Compression (ROHC), detailed in Section 10.17.1, uses window based least significant bits encoding for the compression of dynamic fields in the protocol headers. It is similar to video compression. A base frame and then several difference frames are sent to represent an IP packet flow. As long as the base frame is not lost, it can survive many packet losses in its highest compression state. It is suitable on wireless links with high BER due to its feedback mechanism. It can reduce 40 bytes to 1 byte.

3.9 IP Security

IP security was not imminent in the first era of Internet with time; however, it has become essential to secure the network as well as support the necessary protection for mobile applications as well as virtual private networks. The framework for IP

[9] “In July 2000, Ericsson and Japan Telecom successfully completed a field trial of VoIP over WCDMA using ROCCO...”.

[10] “In May 2002, several companies participated in a successful first trial of the major parts of the ROHC standard including robustness tests over emulated WCDMA/3G links...”.

security (IPSec[11]) has begun in the 1990s by IETF. The IPSec standard is optional for IPv4, however mandatory feature of IPv6 network. An unsecure IP packet is vulnerable to the following; it can be spoofed by eavesdroppers (confidentiality); sender and destination address may be altered along the network (authentication); data can be modified (integrity). These are addressed in IPSec framework by the following services:

- Authentication
- Integrity
- Privacy
- Protection against replays
- Compression

IPSec provides datagram protection with two protocols: authentication header (AH) [RFC2402] and the encapsulating security payload (ESP) [RFC2406] as seen in Figure 3.7:

AH: AH provides data integrity, data source verification, and protection against replay. The host on a secure LAN digitally signs the packet to authenticate it. The receiving host will check the signature and either accept or reject the packet. If the packet has been altered during its journey, the digital signature will not concur with the packet contents. The contents are not encrypted as they travel across the Internet.

ESP: In addition to these, ESP also provides data confidentiality. The host on a secure LAN may encrypt the packet for its journey across the Internet. Anybody that happens listening to packets using a network analyzer will be able to receive the packet but will not be able to decipher its contents since the entire payload is encrypted[12] or authenticated or both.

Key management is not part of the protocol; however, it can be provided manually or through Internet Key Exchange (IKE) protocol [RFC2409], which is based on public-key-based approach for automatic key management. Other automated key distribution techniques such as Kerberos and SKIP may be used as well.

IPSec introduces two modes of operation: *transport mode* and *tunnel mode*. Transport mode is basically with two peers without any intermediate nodes in between. However, tunnel mode is to protect entire IP datagram when packets traverse through security gateways. AH and ESP modes of operation can support transport and tunnel modes. They can be applied individually or in combination with each other to provide a set of security services in IPv4 and IPv6.

In transport mode, AH introduces AH header in between IP header and payload. AH header contains the following fields:

[11] IPSec is defined by several RFCs:

- RFC2401: Security architecture for the Internet protocol
- RFC2402: Authentication header
- RFC2411: IP security document road map
- RFC2406: Encapsulation security payload

[12] “Possible encryption algorithms are DES, 3DES, AES, RC5, IDEA, 3-key triple IDEA, CAST, and Blowfish...”.

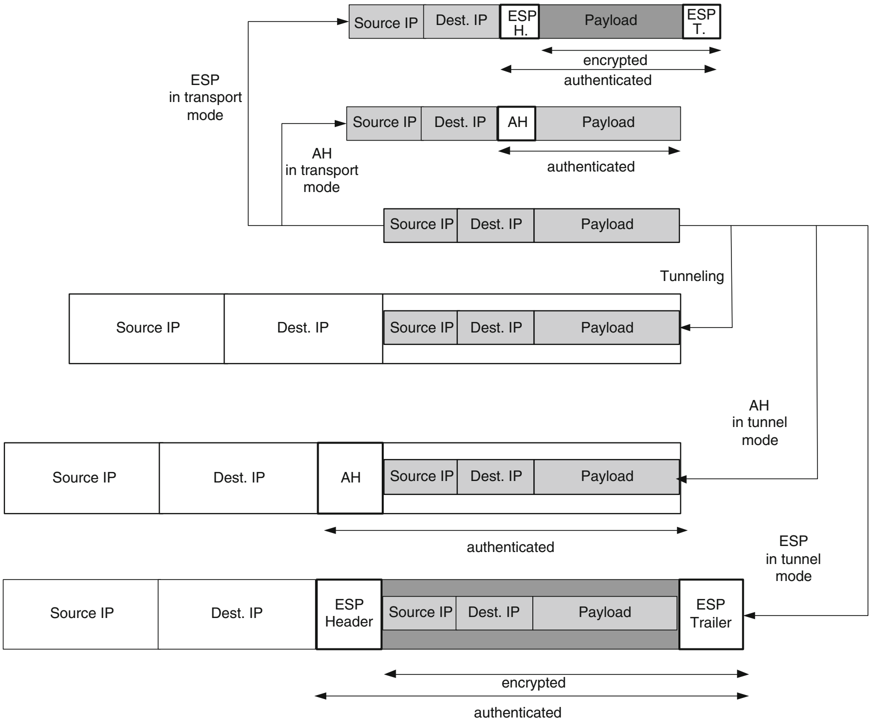


Fig. 3.7 IPSec procedures

- Next header
- Payload length
- SPI
- Sequence number
- Authentication data

Note that IP payload and selected header fields are included in the authentication calculation. However, in tunnel mode, the entire IP datagram (including IP header) is included. Also, tunnel mode requires another IP header in the beginning.

In transport mode ESP, only payload is encrypted; however, in tunnel mode ESP both IP header and payload is encrypted and new unencrypted IP header is placed to the beginning.

3.9.1 Security Associations

IPSec tracks a session through security association (SA). SA defines the structure of the communication system such as authentication algorithm, encryption algorithm also includes information about dataflow, lifetime as well as sequence numbering.

SA is a simplex connection consequently requires two SA for each peer; that is, one from X to Y and one from Y to X. A SA is uniquely identified by the following:

- Security parameter index (SPI). This in turn can be associated with key data, encryption and authentication algorithms
- IP destination address
- A security protocol (AH or ESP) identifier

Information about security association is stored in security association database (SADB) and also there is security policy database (SDP) to define which traffic security policy to be applied to which destination. The setting up of a SA between two peers on the network is not defined within IPSec and can be performed in number of different ways.

IPSec works in parallel with a key management mechanism, which is required to negotiate parameters for each SA. The IKE (new version IKE (SOI) or IKEv2) is one of the popular methods to establish keys and SAs. IKE creates a SA for a communication and define encryption, authentication, and integrity algorithms in addition to the keys. IKE protocol first establishes a shared secret between peers and ensures that the communication is with the right peer. Then, IKE authenticates the peer and creates a pair of SA.

3.10 IP Tunneling

All-IP network will increase the number of independent networks a packet traverse to reach the destination. To establish a one hop communication behavior, tunneling is introduced. Tunneling is a method of using an internetwork infrastructure to transfer data of one network over another network. Instead of sending the frame as it is produced by the source, the tunneling protocol encapsulates the frame in an additional header as seen in Fig. 3.8. The new header enables original frame to

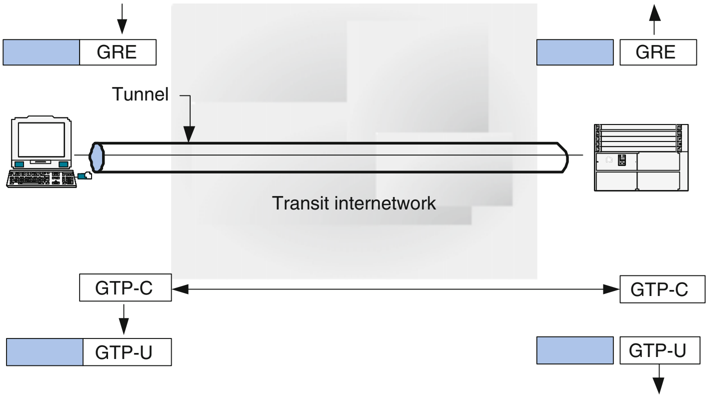


Fig. 3.8 Tunneling

be carried over the intermediate network. The encapsulated packets are then routed between tunnel endpoints over **the tunnel**. In the destination, the frame is decapsulated and forwarded to the final destination. Hereby, tunneling includes this entire process: encapsulation, transmission, and decapsulation.

Tunneling technology can be based on Layer 2 or a Layer 3. Point to Point Tunneling Protocol (PPTP), Layer Two Tunnel Protocol (L2TP) are some well known examples for Layer 2 tunneling protocols. IPSec tunnel, Generic Routing Encapsulation (GRE), and GPRS Tunneling Protocol (GTP) are Layer 3 protocols. GRE and GTP tunneling protocols are selected as tunneling protocol in WiMAX and LTE, respectively.

In brief, PPTP [RFC2637] is designed to enable secure transfer of data from a remote client to a private enterprise server by creating a Virtual Private Networking (VPN) across TCP/IP. PPTP establishes a regular PPP session via GRE protocol. There is another session over TCP to initiate and manage the GRE session. PPTP are authenticated with EAP-TLS or MSCHAP-v2. L2TP [RFC2661, version 3 RFC3931] combines the best features of Microsoft's PPTP and Cisco's Layer 2 Forwarding (L2F). L2TP uses UDP to transmit its payload and is transparent to the higher level protocols. Tunnel is established between LAC (L2TP Access Concentrator) and LNS (L2TP Network Server) and once a tunnel is established, the network traffic is bidirectional.

GRE[13] designed by CISCO is a *stateless* protocol where the tunnel end-points do not monitor the state or availability of other tunnel end-points. GRE does not provide any encryption by itself but provides an efficient low overhead tunneling. GRE is often used in conjunction with network layer encryption protocols such as IPSec to provide encryption. GRE tunneling uses the entire packet as the payload and appends a GRE header. Forwarding GRE encapsulated packets will not be any different than other packets in any way. When a GRE packet is received, the receiver first needs to interpret that it is a GRE packet. Once this is determined, the receiver after verifying the GRE packet with information contained in the header removes the header.

GTP may run over TCP or UDP; however, it requires explicit signaling to set up tunnels. Initially, in GTP version 0 (GTPv0), the signaling protocol that sets up the tunnels is combined with the tunneling protocol on one port. This is split into two with GTP version 1 (GTPv1), which introduces two protocols one for control (GTP-C) and one for user data tunneling (GTP-U). There is also GTP' for accounting. Lately, 3GPP has introduced GTP version 2 [TS 29.274] (aka evolved GTP or eGTP) for LTE (Evolved Packet Systems). GTPv2 also has two protocols: control plane protocol (GTPv2-C) and the user plane protocol (GTPv2-U). A GTP tunnel is identified in each node with a TEID (Tunnel Endpoint Identifier)

[13] GRE is defined by several RFCs:

- RFC 1701: Generic Routing Encapsulation (GRE).
- RFC 1702: Generic Routing Encapsulation over IPv4 networks.
- RFC 2784: Generic Routing Encapsulation (GRE).
- RFC 2890: Key and Sequence Number Extensions to GRE.

and a UDP port number. The receiving end point assigns the TEID for the transmitter to use and TEID values are exchanged via GTP-C signalling as seen in Fig. 3.8.

3.11 PPP: Point-to-Point Protocol

PPP is an encapsulation protocol to transmit IP traffic between two peers. It has emerged as a Layer 2 protocol over synchronous modem links, as a replacement for the non-standard Layer 2 protocol SLIP (Serial Line Internet Protocol). However, other protocols other than IP can also be carried over PPP, including DECnet and Novell's Internetwork Packet Exchange (IPX). Two common encapsulated forms of PPP, Point-to-Point Protocol over Ethernet (PPPoE) or Point-to-Point Protocol over ATM (PPPoA), are used in DSL. PPP negotiates configuration parameters at the beginning of the connection, and these details are remain transparent to the user. PPP is comprised of the following main components:

- A derivative of the High-Level Data Link Control (HDLC) protocol for encapsulating datagrams over serial links. The PPP encapsulation allows multiplexing of different network-layer protocols simultaneously over the same link.
- A Link Control Protocol (LCP) to establish, configure and test the data link connection.
- A group of network control protocols (NCPs) for establishing different network layer protocols.

3.11.1 LCP Link Establishment

To establish a link, first initiator sends a Configure-Request signal. This can contain a number of requested link options, for example the maximum frame size or the authentication protocol for the link. The responder responds with one of the following:

- Configure-ACK: acceptance notification of the request
- Configure-NAK: reject with some options to renegotiate
- Configure-Reject: reject the options requested by sender

The link stays active until explicit LCP or NCP packets close the link down, or until some external event such as inactivity timer expires or network administrator intervention occurs. To close the link, Terminate-Request is sent followed by a response of the Terminate-ACK. The LCP is used as a control protocol to

- agree upon the encapsulation format options
- handle varying packet size limits
- detect a looped-back link
- handle misconfiguration errors

- provide authentication of the identity of its peer
- determine the proper functioning of the link
- terminate link

After the link has been established and optional facilities have been negotiated, PPP must send NCP packets to choose and configure one or more network-layer protocols. Once they have been configured, datagrams of each network-layer protocol can be sent over the link. The NCP protocol for IP is called PPP Internet Protocol Control Protocol or IPCP [RFC1332]. IPCP allows the configuration of IP options such as the IP host address and header compression.

3.11.2 PPP Authentication

PPP also offers the following methods for authentication: Password Authentication Protocol (PAP), Challenge-Handshake Authentication Protocol (CHAP), and Microsoft Challenge Handshake Authentication Protocol (MS-CHAP). PAP is a simple clear-text authentication scheme. It is insecure since it sends the user's login name and password unencrypted across the channel. CHAP is more secure since the login name and password are encrypted before being sent across the channel [RFC1999].

Let us say User A wants to authenticate. User A sends a long random number as a "challenge" to the authenticator. Authenticator responds with a value calculated using a one-way hash function which takes the random number and shared secret (user's password) as inputs. If User A sees that the hash value is correct, User A is authenticated and success is indicated, otherwise a failure message is returned. If the random numbers are very long, then hacker can not fool the authentication mechanism by recording the responses. Also password is not sent over the network. RFC1999 allows for various hash algorithms to be used but only requires support for MD5. The length of the response values for MD5 is 16 bytes. Identity of the system is indicated by the name on the challenge and response packets. CHAP requires that both the user and authenticator know the shared secret, although it is never sent over the network. This requirement is eliminated with MS-CHAP, which does not require either peer to know the shared secret.

3.12 AAA

AAA (Authentication, Authorization and Accounting) framework is standardized by IETF to address the need for remote access. The AAA framework is important for mobile users who access Internet through an access network such as WiMAX, LTE, WiFi, etc. As a result, identity of a remote user first needs to be validated by authentication, and then user needs to be authorized for a given service in a network. This is also complemented with accounting to facilitate billing on the extent of service usage.

The AAA has introduced Remote Authentication *Dial-In* User Service (RADIUS) [RFC2865] to carry out authorization, authentication, and accounting functions between the access server and the user information repository. Note that AAA framework has started in dial-up era and extended to mobile networking. IETF AAA working group has introduced new framework called DIAMETER[14] for the next generation AAA server with backward compatibility to RADIUS to ease migration. Mobile IP ROAMOPS (Roaming Operations) TR45.6 working group is defining the requirements for DIAMETER. A primary difference between DIAMETER and RADIUS is that DIAMETER allows peers to exchange a variety of messages.

The AAA framework is established between AAA server, which has AAA information about each user and AAA client located in Network Access Server (NAS). A user communicates with NAS to request access to the network. NAS is responsible to communicate to AAA server. The AAA server processes the data and sends acceptance or rejection to the AAA client. Also AAA server informs AAA client about authorization and accounting information upon acceptance. If a user is connected to another ISP other than its home ISP, home network location for AAA server is found by using a network address identifier (NAI), which is basically like an email address (name@home.network).

3.12.1 RADIUS

The RADIUS protocol is widely used today. RADIUS is based on **client/server** model where communication in between is authenticated through the use of a shared secret. RADIUS client accesses a centrally located RADIUS AAA server over UDP to validate user identity and retrieve all configuration information. For instance, in WiMAX, RADIUS client resides in the ASN-GW as seen in Fig. 3.9 and RADIUS AAA performs the following functions:

- *Authentication* determines the identity of a user via an encrypted key. The RADIUS may support a variety of methods to authenticate a user. It can support PPP with PAP or CHAP, login screen, and other authentication mechanisms.

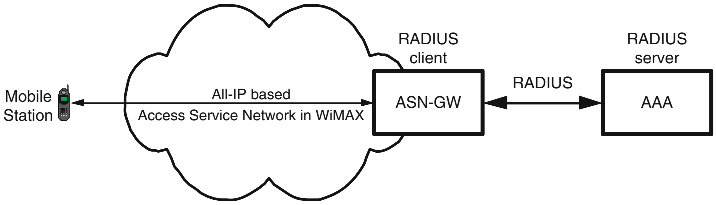


Fig. 3.9 RADIUS configuration

[14] According to the DIAMETER RFC: "The basic concept behind DIAMETER is to provide a base protocol that can be extended in order to provide AAA services to new access technologies. Currently, the protocol only concerns itself with Internet access, both in the traditional PPP sense as well as taking into account the ROAMOPS model, and Mobile-IP."

- *Authorization* allows or rejects user access to the network based on their profile and the current security policy.
- *Accounting* controls usage statistics for accounting purposes [RFC2866]. RADIUS client sends accounting-request to turn *on*, start, stop, and turn *off* accounting and RADIUS server responds with accounting-response.

RADIUS is vulnerable if shared secret between client and server is known. This may enable intruders act as client or server to collect user information. Also first message (access request) is not authenticated, which may increase the insecurity. Another drawback is its limitation to allow unsolicited messaging. Hence, new service deployment is restricted.

3.12.2 DIAMETER

DIAMETER, which addresses the flaws of RADIUS, is composed of a base protocol and a set of protocol extensions. Base protocol provides the basic functionality that is common to all services supported in DIAMETER; however, application specific functionalities are provided through extension mechanisms. DIAMETER base protocol creates relationship between Mobile IP Application, NAS Application, and SIP (Session Initiation Protocol) Application with support for TLS (Transport Layer Security) and IPSec. DIAMETER can also be used to provide Extensible Authentication Protocol (EAP) to PPP users and transport X.509 digital certificates.[15] DIAMETER is currently the core protocol of IP Multimedia Subsystem (IMS) architecture both in service and control plane along with SIP (Session Initiation Protocol).

[15] "ITU-T X.509v3 certificate [RFC 2459] consists of three parts:

- A certificate body containing
  - a version number (currently v3, v2 and v1 are also possible)
  - a unique serial number assigned by the responsible certificate authority (CA)
  - a declaration of the signature algorithm to be used to sign the certificate such as md5RSA
  - the ID of the CA that issued and signed the certificate
  - the validity period (not valid before/not valid after)
  - the subject (user) ID
  - the public key of the subject (user)
  - any number of optional v2 or v3 extensions, some of them being very important
- A definition of the signature algorithm used by the CA to sign the certificate
- The signature guaranteeing the authenticity of the certificate, consisting of the hashed certificate body encrypted by the CA's private key

Note that to verify a certificate, you need to verify the CA that issues the certificate. Also, the CA has a certificate that is issued by another CA. X.509 forms this hierarchical CA structure where there is a *Root* CA at the top of the chain. Examples of well-known Root CAs are Verisign, RSA, Baltimore, Entrust, Thawte, Deutsche Telekom and Swisskey. As a result, it is ample to have at least one trust relationship with any CA in the chain to verify the certificate."

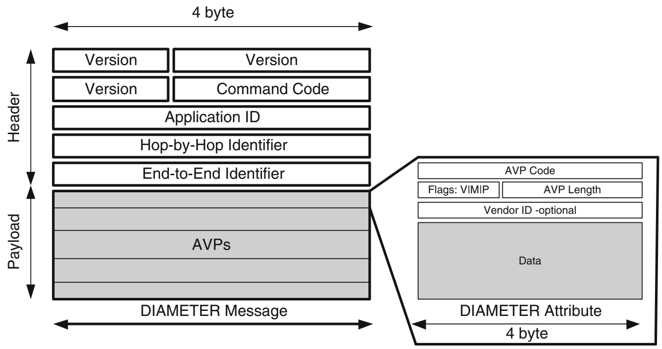


Fig. 3.10 DIAMETER packet format

Unlike RADIUS which is a client/server protocol, DIAMETER is a *peer-to-peer protocol*; a node can be a client, server, or an agent. There are relay, proxy, redirect, and translation agents; *relay* agent forwards a message to the appropriate destination and can be utilized as a aggregation point to reduce the overhead; *proxy* agent additionally has a right to modify the messages; *redirect* agent upon receiving a message checks its routing table and returns a response message with redirection information to its original sender; *translation* agent converts a message from AAA protocol to another to ease migration. This enables service providers to support mobility between many different domains.

DIAMETER packet format is depicted in Fig. 3.10, which predefines a set of Attribute-Value-Pairs (AVPs) to carry the details of AAA, routing, security:

- Abort-Session Request/Response
- Accounting Request/Response
- Capabilities-Exchanging-Request/Answer
- Device-Watchdog-Request/Answer
- Disconnect-Peer-Request/Answer
- Re-Auth-Request/Answer
- Session-Termination-Request/Answer

DIAMETER also provides peer discovery to avoid manual configuration of the NAS with the location of its DIAMETER AAA server. DIAMETER server or agent broadcasts which application they support, along with provided security level. Then, DIAMETER client finds suitable first-hop to forward DIAMETER messages. After selecting the peer, connection is established with TCP or SCTP.

3.13 EAP: Extensible Authentication Protocol

IEEE has introduced IEEE 802.1x to provide authentication and authorization of devices attached to a network. There is also Extensible Authentication Protocol (EAP) [RFC2284, now RFC3748] that is used to exchange authentication messages. IEEE

802.1x is first defined for wired networks then extended to provide authentication and authorization services to IEEE 802.16 and IEEE 802.11.

IEEE 802.1x has introduced three entities: *supplicant* (user device), *authenticator* (ASN-GW in WiMAX), *authentication server* (RADIUS server). The goal of 802.1x/EAP is to distribute the shared key (master key) between the supplicant and the authentication server. An EAP packet has four fields: code, identifier, length, and data. EAP constitutes bidirectional communication between supplicant and authentication server. EAP security varies depending on the EAP method selected: EAP-TLS, PEAP, EAP-TTLS, EAP-FAST [RFC4851], EAP-SIM, EAP-AKA, EAP-MD5, EAP-PSK, EAP-IKEv2, etc. We herein introduce three EAP methods in detail.

3.13.1 EAP-TLS

EAP-TLS [RFC5216] is SSL/TSL procedure carried over EAP packet as seen in Fig. 3.11. TLS is a client/server application to prevent eavesdropping, tampering, or message forgery. The TLS introduces a handshake protocol which enables peers to agree upon security parameters to authenticate themselves and determine keys for data encryption and integrity protection. As can be seen in Fig. 3.11, ClientHello and ServerHello messages are responsible to determine the algorithm to use and exchange random values. TLS uses public key cryptography algorithm to calculate the shared keys. Before this, peer certificates should be verified. EAP-TLS provides

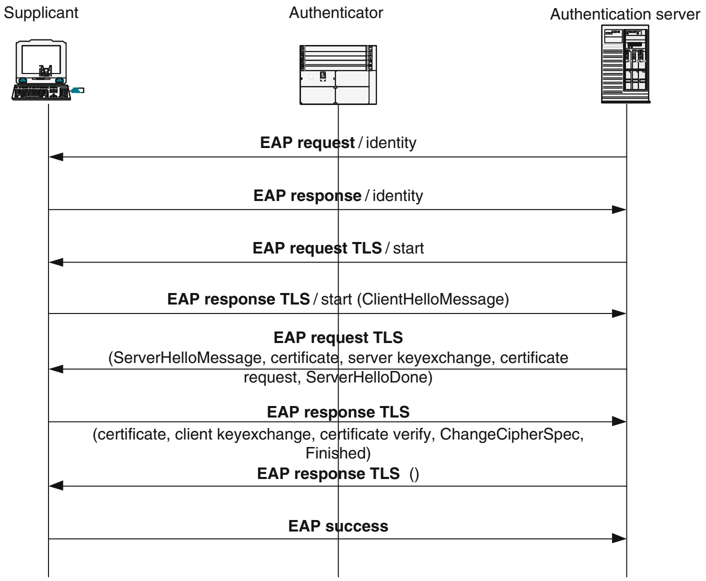

Supplicant
Authenticator
Authentication server
EAP request / identity
EAP response / identity
EAP request TLS / start
EAP response TLS / start (ClientHelloMessage)
EAP request TLS
(ServerHelloMessage, certificate, server keyexchange, certificate request, ServerHelloDone)
EAP response TLS
(certificate, client keyexchange, certificate verify, ChangeCipherSpec, Finished)
EAP response TLS ()
EAP success

Fig. 3.11 EAP-TLS

excellent security but overhead is introduced with client-side certificates. Client-side certificates can be housed in smartcards to achieve highest level of security. EAP-TLS provides mutual authentication and key generation; however, it lacks identity protection and unprotected EAP success and failure messages. This is addressed in Protected EAP (PEAP) and EAP-TTLS.

3.13.2 EAP-TTLS

EAP-TTLS employs *tunneled TLS* and removes the need of client authentication but server authentication to client is required. EAP-TTLS introduces a TTLS server which establishes a secure tunnel to authenticate the client. The tunnel provides protection from eavesdropping and man-in-the-middle attack. TTLS server is responsible to transmit AVPs sent by client to the AAA server. EAP-TTLS also provides supplicant identity protection and data ciphering suite negotiation in addition to the features offered by EAP-TLS.

3.13.3 EAP-AKA

EAP-AKA [RFC4187] is considered in 3GPP for UMTS Authentication and Key Agreement using the Universal Subscriber Identity Module (USIM). Its predecessor is EAP-SIM [RFC4186], which is used for GSM authentication and key distribution using Subscriber Identity Module (SIM). EAP-SIM only protects security between mobile station and base station since network beyond base station is assumed to be secure. EAP-AKA extends this beyond base station and provides no area with clear data transmission. In 3G, AKA is used both for network access as well as IMS. Different user identities and formats are available; International Mobile Subscriber Identifier (IMSI) which is a 15 digit number is used for network access, whereas the IMS uses the Network Access Identifier (NAI). Figure 3.12 illustrates the EAP-AKA procedure in which EAP master key is generated by the supplicant and the authenticator. On the basis of NAI, appropriate 3GPP AAA server is selected, which retrieves the subscriber's profile and authentication vector AV(1..n), based on the secret key and a sequence number from Home Subscriber Server (HSS). 3GPP AAA server stores the authentication vector: XRES (expected response), AUTN (authentication value), CK (ciphering key), and IK (integrity key). Then, AAA server transmits AKA-challenge including RAND (random number), AUTN to the USIM in supplicant. The USIM calculates the RES, CK, IK, and sequence number (SQN). SQN is maintained at HSS and USIM and not revealed to the network. USIM transmits RES (result) to 3GPP AAA server for comparison with XRES – to authenticate itself – but also checks the computed value of the SQN with its own version to authenticate the network to itself.

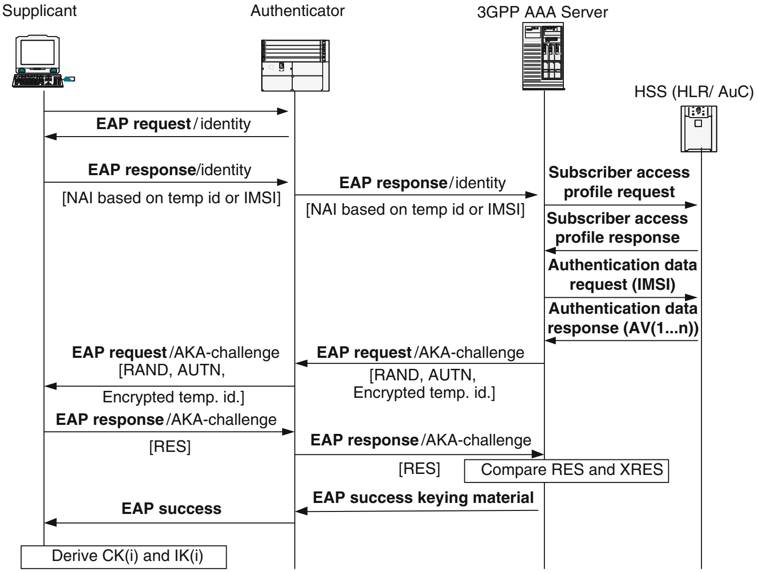

Supplicant
Authenticator
3GPP AAA Server
HSS (HLR/ AuC)
EAP request/identity
EAP response/identity
[NAI based on temp id or IMSI]
EAP response/identity
[NAI based on temp id or IMSI]
Subscriber access profile request
Subscriber access profile response
Authentication data request (IMSI)
Authentication data response (AV(1...n))
EAP request/AKA-challenge
[RAND, AUTN,
Encrypted temp. id.]
EAP request/AKA-challenge
[RAND, AUTN,
Encrypted temp. id.]
EAP response/AKA-challenge
[RES]
EAP response/AKA-challenge
[RES]
Compare RES and XRES
EAP success
EAP success keying material
Derive CK(i) and IK(i)

Fig. 3.12 EAP-AKA

IK is used to authenticate the signaling messages between mobile and network and CK is used to encrypt the data over the air. Also IMSI encryption is performed at first connection with a group key and P-TMSI (Packet Temporary Mobile Subscriber Identifier) is first used to prevent IMSI being captured for malicious use and impersonation of users.

3.14 Mobile IP

Mobile IP[16] is designed by Internet Engineering Task Force (IETF) to allow mobile subscribers to move from one network to another while maintaining a permanent IP address to maintain the usage of resources and services on the move.

The packets in Internet are routed to the network the subscriber's IP address belongs. To maintain the seamless roaming, Home Agent (HA) entity is introduced in the home network. Mobile subscriber (MS), when leaves the network, initiates a binding request to its Home Agent via Foreign Agent (FA) residing in the visited network. Foreign Agent advertises a care-of address (CoA) in the visited network and links the Mobile subscriber's home address (HoA) to the CoA by sending a binding request.

[16] Mobile IPv4 is described in IETF RFC3344, and updates are added in IETF RFC4721. Mobile IPv6 and IPSec to protect Mobile IPv6 are described in IETF RFC3775 and IETF RFC3776 respectively.

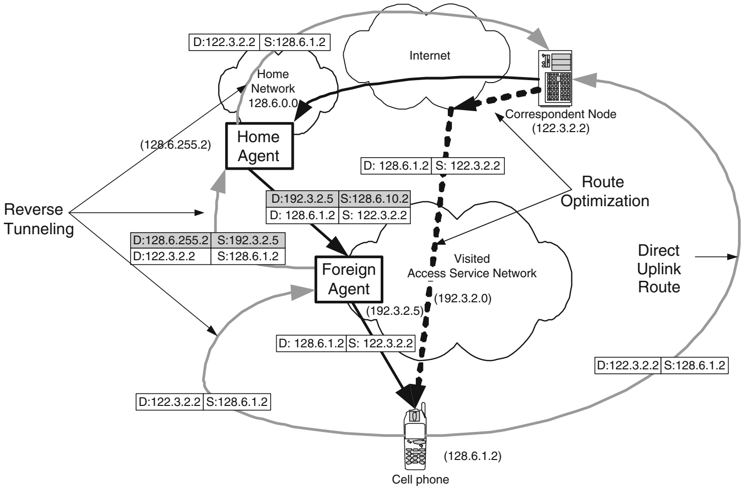


Fig. 3.13 An mobile IP

Figure 3.13 illustrates a Mobile IP procedure. Packets are sent from a correspondent node (CN) to the destination network. The HA intercepts these and diverts them to the FA using the mobile node's care-of address. HA encapsulates the packets with an outer IP header and sends to FA through a tunneling protocol. FA, when receives the tunneled packet, removes the outer header and forwards the contents to the mobile user directly. Security to prevent denial of service attack on the network is established with a shared key between MS and HA. HA checks the key every time there is a request.

3.14.1 Route Optimization

The route can be optimized as illustrated in Fig. 3.13 by making aware of the CN about the care-of address. But it is difficult for security reasons, authentication between the MS and the HA is relatively simple since they can both be configured with a shared secret. Authentication between the correspondent node and the MS is a lot more difficult, since the correspondent can be any node on the Internet. As a result, it might not be feasible to authenticate MS with every node in the Internet.

3.14.2 Reverse Tunneling

In the upstream direction if MS sends packets directly to the correspondent node, there is a mismatch: if we follow the Fig. 3.14, the network prefix for these packets

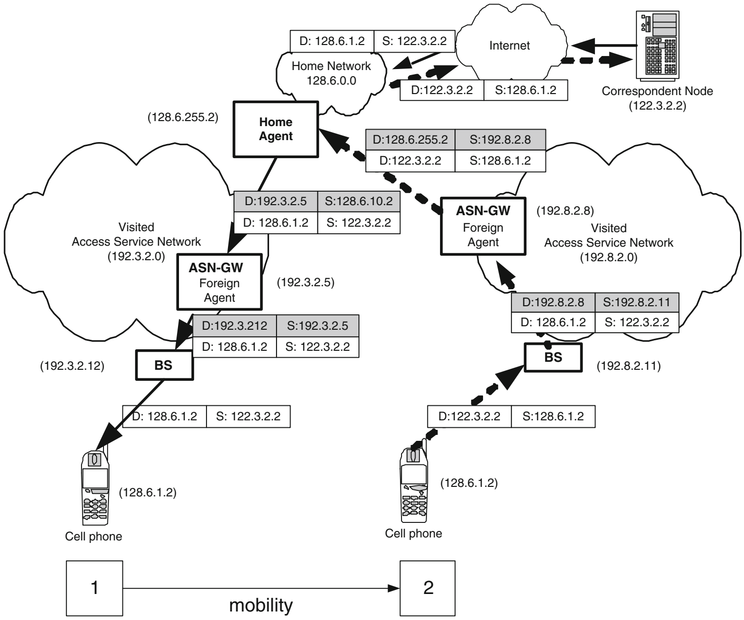


Fig. 3.14 An example for mobile IP procedure in WiMAX

are 128.6, that of its home network. However, it is now residing in network with prefix 192.8.2. Security devices (e.g., firewalls) may filter out these packets since it may mark them as illegal IP source addresses. This is to protect network to some types of denial of service attacks. Reverse tunneling addresses this issue by reversing the outgoing transmission as in the incoming route. MS sends the packets to FA, FA tunnels them to HA, and HA removes the tunnel and forwards the packet to the final destination.

3.14.3 PMIPv4: Proxy Mobile IPv4

So far we assume a Mobile IP stack in the subscriber to perform the signalling. This mode of operation is called Client Mobile IP (CMIP). There is also another type of Mobile IP called Proxy Mobile IP (PMIP), which a network-based mobility and the mobile subscriber is not involve in the mobility signaling instead there is a PMIP agent in the network that performs signalling on behalf of the subscriber. Figure 3.14 illustrates the PMIPv4 implementation in WiMAX. PMIP client resides in ASN-GW with FA and is responsible to perform the binding. Also note that there is additional tunnel between ASN-GW and BS since they reside in different networks. Mobile subscriber only knows and maintains one IP address, which is assigned by the network during registration.

3.14.4 Mobile IP for IPv6

Mobile IP with IPv6 is simpler and more scalable than that with IPv4. It uses the inherent security mechanisms provided by IPv6 (i.e. IPSec) and supports route optimization. Goals of IPv6 mobility includes the following:

- Always-on IP connectivity
- Session persistence
- Static IP addresses
- Roaming between L2 and L3 networks

For IPv6, there is no use of FA. The CoA is directly colocated with MS. When MS connects to a visited network, it gets a CoA and sends binding request to HA. The binding request is secured with IPSec ESP in transport mode. HA encapsulates every traffic with IPv6-in-IPv6 tunnel and sends it to the CoA of the MS. From MS, the packets are reverse tunneled CN via HA.

In IPv6, route optimization is possible, MS sends binding update to CN with CoA as source address. This bypasses the anti spoofing in the visited network. CN replaces the source address with the HoA of MS and passes the packet to the upper layer protocols. In the downlink direction, CN sends traffic to MS with CoA as destination address with a special routing header with HoA as second hop. MS removes the routing header and forwards the packet to the upper layers.

Security of binding is established between MS and HA with trust relationship via IPSec with ESP in transport mode. Between MS and CN, trust is established with Return Routeability procedure.

3.14.4.1 Return Routeability Procedure

MS sends two cookies to CN in two different paths. MS sends Home Test init (HTi) cookie via HA and Care-of Test init (CTi) cookie directly to CN. CN builds two key generation tokens (home keygen and care-of keygen) with a random key and nonce. CN sends back this key generation tokens and cookies to MS directly and via HA. MS builds a binding message key with these two key generation tokens and secures the binding request message. This helps CN proof that MS is reachable via both paths.

3.14.5 PMIPv6: Proxy Mobile IPv6

Mobility for IPv6 nodes without a MIPv6 stack is possible by a proxy mobile agent in the network, which performs the messages with home agent. This protocol is referred to as PMIPv6. The PMIPv6 introduces Local Mobility Anchor (LMA) and Mobile Access Gateway (MAG). LMA acts as a home agent within the PMIPv6 domain, which is defined as the network in which mobility is handled using the PMIPv6 protocol. LMA is responsible to maintain the subscriber's reachability state and is the anchor point for subscriber's home network prefix(es). MAG is the mobility management entity and performs the necessary mobility related signaling for a mobile subscriber.

The mobile subscriber may be an IPv4 node or IPv6 node or dual stack node. Depending on the configuration MS receives HoA (IPv4, IPv6, or dual IPv4/IPv6 addresses) and moves in the PMIPv6 domain. After MS is authenticated, the AAA delivers LMA address (LMAA) of the MS to the MAG. The MAG sends Proxy Binding Update (PBU) message with its address, Proxy-CoA, and the MS address to the LMA. The LMA registers the MAG for the MS and sends Proxy Binding Acknowledgement (PBA) message which includes Home Network Prefix of the MS to the MAG. With PBA message, MAG establishes a bidirectional tunnel between LMA and MAG where LMAA and Proxy-CoA identifies the tunnel endpoint. The MAG sends Router Advertisement (RA) message including the home network prefix to the MS. The MS configures its address (MN-HoA) with this home network prefix and uses MN-HoA to send and receive data.

Any packet outside PMIPv6 domain will be received by the LMA. The LMA forwards these to MAG via bidirectional tunnel. The MAG removes the tunnel header and forwards the packet to the MS. If CN is locally connected to MAG then packets may be routed locally by the MAG. In the upstream direction, any packet sent my MS is received by the MAG, which tunnels these to LMA. The LMA strips of the tunnel header and forwards the packet toward destination.

3.15 SIP: Session Initiated Protocol

SIP[17] [RFC3261] is a control signaling protocol residing in application-layer. It is transport-independent and can be over TCP, UDP, ATM, or SCTP. SIP is mainly used for multiuser involved applications such as conference calls, instant messaging, multimedia calls, etc. SIP is designed as an IP protocol and suits good to All-IP Networking. SIP users have SIP uniform resource indicator (URIs) like

sip:xyx@example.com
sip:+15101234567@example.com;user=phone

where first one is for user who is connected to Internet and second one is for user from PSTN.

SIP components depicted in Fig. 3.15 are as follows:

- **SIP Agent service:** User Agent (UA) is the peer entity. Between two user agents a call is set up. UA can be subscriber or an application in the server. For instance, a video streaming would have a subscriber and application as UAs.

[17] "It was originally designed by Henning Schulzrinne (Columbia University) and Mark Handley (UCL) starting in 1996. The latest version of the specification is RFC3261 from the IETF SIP Working Group. In November 2000, SIP was accepted as a 3GPP signaling protocol and permanent element of the IMS architecture. It is widely used as a signaling protocol for Voice over IP, along with H.323 and others...".

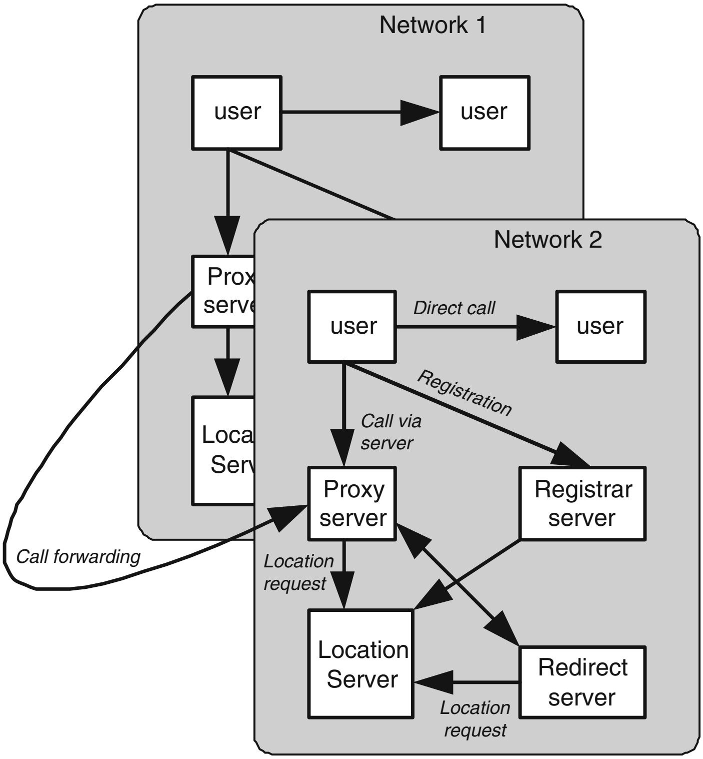


Fig. 3.15 SIP components

- **SIP Proxy service:** Proxy server forwards requests to the correct location of the UAs.
- **SIP Registrar service:** UAs register their current location info to Registrar server. Location info may include IP address, cell ID, and the ID of P-CSCF.
- **SIP Redirect service:** UA request is responded with redirection response for UA to initiate new call to the location.
- **SIP Location service:** UA location info is stored and updated.

When SIP is used, certain QoS requirements are provided via SDP content. Proxy server determines the bandwidth and requests the QoS according to SDP content. Figure 3.16 illustrates a call setup with SIP proxy. SIP uses few basic message types with growing number of extensions:

- INVITE is used to initiate basic SIP call.
- ACK is used to acknowledge the INVITE.
- BYE is used to terminate a call.
- CANCEL is used to terminate a call before being established.
- REGISTER is used to indicate UA their location and availability.
- OPTIONS is used to determine the capabilities.

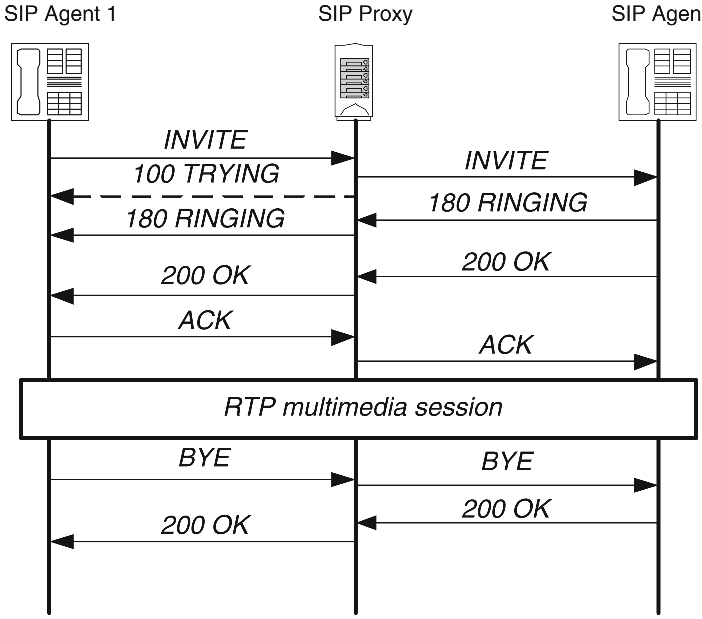


Fig. 3.16 SIP call with SIP Proxy: Notice that proxy can be stateless and if stateless proxy is used then there is no 100 TRYING message

- INFO is used to carry signaling during SIP call.
- PRACK is used to carry reliable delivery of provisional responses.
- UPDATE is used to modify the media information in terms of QoS before full establishment.
- REFER is used to establish a connection with SIP agent and a third party agent.
- SUBSCRIBE is used to subscribe to an event notification.
- NOTIFY is used to notify the agent when message arrives.
- MESSAGE is used to support instant messaging.

SIP entity first establishes a connection to proxy or registrar service by getting the IP address/port tuple from DHCP and DNS. SIP entity has a set of response codes for each request. There are six type of classes (1xx, 2xx, 3xx, 4xx, 5xx, 6xx). For instance, 100 TRYING and 180 RINGING in Fig. 3.16 belong to 1xx class, which is used for provisional/informational responses. 2xx is used for success as in 200 OK in the figure, 3xx is used for redirection, and 4xx/5xx/6xx are used for client/server/global errors, respectively. SIP transactions are reliably handled with request and response signal exchanges. If response does not arrive within a time interval, retransmission is performed and timeout period is doubled.

Calls within IP network is handled in SIP with DNS and IP routing but calls from IP to PSTN are handled by assigning the SIP URI to destination number. When calls are from PSTN to IP, then E.164[18] numbers (ENUM) are mapped to IP addresses in the PSTN/SIP gateway [RFC2916] by reversing the digits (ex: +90 (312) 231 7975) and appending *e164.arpa* suffix for DNS address as:

5.7.9.7.1.3.2.2.1.3.0.9.e164.arpa

By this, each user in the IP network holding a unique E.164 number can be addressable through DNS server.

3.16 IMS: IP Multimedia Subsystem

IMS is an open, standardized multimedia architecture for mobile and fixed IP services originally defined by the Third Generation Partnership Project (3GPP), and largely adopted by the Third Generation Partnership Project 2 (3GPP2). The IMS is based on SIP, DIAMETER, and RTP protocols, standardized by IETF. The key feature is being access agnostic in which IMS services can be retrieved from any location with flexible charging models.

IMS started as a technology for 3G mobile networks, but it is now spreading to next-generation networks. IMS builds on Session Initiation Protocol (SIP), which has emerged as the crucial technology for controlling communications in all-IP-based networks. IMS service offers quality in user experience during a change in the domain. Thus, IMS may support many standards simultaneously and dynamically.[19]

IMS[20] aims to offer rich multimedia services across both next-generation packet-switched and traditional circuit-switched networks with standards-based open interfaces and functional components. Network owners can simultaneously derive added value from their networks and open these networks to third parties to develop and offer services and applications of their own.

[18] "E.164, an ITU-T recommendation, defines the international public telecommunication numbering plan used in the PSTN and some other data networks and restricted to maximum of 15 digits."

[19]

- WiMAX
- 3GPP IMS
- 3GPP2 Multimedia Domain
- TISPAN (Telecoms & Internet converged Services & Protocols for Advanced Networks)
- ITU-T FG NGN (International Telecommunications Union standard for next-generation networks)
- ATIS NGN-FG (Alliance for Telecommunications Industry Solutions Next Generation Networks - Functional Group)
- IETF (Internet Engineering Task Force)
- CableLabs' PacketCable 2.0

[20] "...Softbank Mobile Corp in Japan launched the world's first IMS-based services in November 2006 over a 3G network with new exciting 3G services initially including push-to-talk, presence and group list management. IMS Mobile VoIP over HSPA was demonstrated for the first time on a mobile terminal at the World Congress 2007..." Source: `wwww.3gamericas.org`.

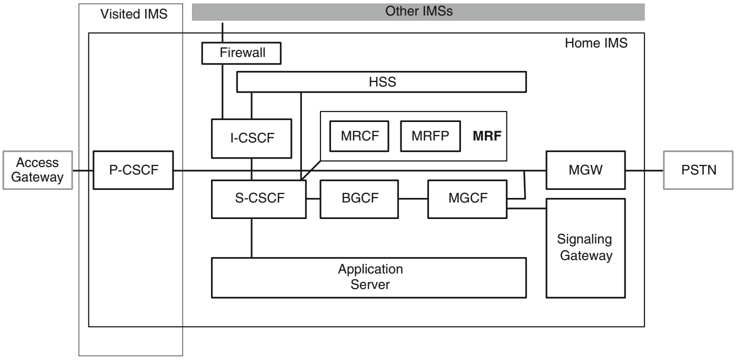


Fig. 3.17 IMS functional decomposition

Figure 3.17 shows the IMS architectural components and interfaces. IMS can be connected to other operator's IMSs to maintain continuity of services. IMS is connected to mobile network through access gateway and circuit-switched network through ISDN. The components in Fig. 3.17 are defined as follows:

- **P-CSCF:** Proxy-Call Session Control Function is the first entity a user connects to either in home IMS or in visited IMS. P-CSCF forwards the call signaling to the serving CSCF.
- **S-CSCF:** Serving-Call Session Control Function is responsible to carry out session and accounting control. Each user is served only by home S-CSCF to ease charging.
- **I-CSCF:** Interrogating-Call Session Control Function is located at the edge for SIP signaling coming from other networks. Either there is a registration request in which it is routed to S-CSCF or a call setup request for user.
- **AS:** Application Server hosts services for subscriber. Services include but not limited to streaming video, multimedia mail, VoIP, etc.
- **HSS:** Home Subscriber Server is the database of user information. Database information includes ID information, security information, location information, and profile information. HSS should be accessible by all the policy controllers in the network so subscriber data are available to all applications in real time. This way subscriber benefits from single sign-on, centralized authorization of identity and unified subscriber self management and billing. The HSS is visible to other applications with Sh and IMS Service Control.
- **BGCF:** Breakout Gateway Control Function is used to select the circuit-switched network to forward the calls coming from S-CSCF.
- **MGCF:** Media Gateway Control Function controls Media Gateway (MGW) for connection between the IMS and circuit switched network. External signaling (SS7/ISUP) from circuit-switched network is translated to SIP in MGCF.
- **MRF:** Multimedia Resource Function is an optional entity for multimedia conferencing. It has control and processing components.

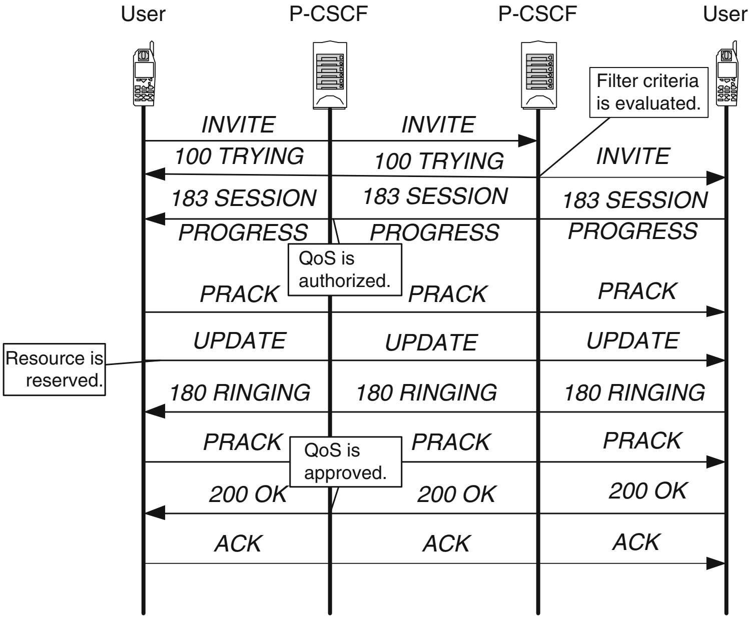


Fig. 3.18 IMS call flow

The signaling protocol on IMS is based on SIP. The CSCF components act as a SIP server and assists accounting. P-CSCF and I-CSCF are proxy servers. S-CSCF holds proxy and registrar server to perform authentication and call routing to appropriate P-CSCF for IMS-to-IMS calls, to BGCF for IMS-to-PSTN calls or to Internet for public calls. Figure 3.18 shows a call flow example for IMS.

New features are being designed in the context of 3GPP and other technologies.[21] Common IMS provides a solution to open 3GPP's definition of IMS to all access technologies. Multimedia Priority Service enhances the IMS to provide special care during disaster recovery and national emergency situations. This gives higher priority to users with suitable authorization during network overload situation. This might require that MPS members are from government or emergency services. IMS is also considering to support packet cable community to maximize the commonality of IMS specification. IMS Service Brokering is another enhancement to enable proliferation of new services based on IMS capabilities. This framework intends to provide simpler development interface and facilitates its interactions with the developed services. There are two other areas to facilitate the communication toward packet services; IMS Centralized Services make sure that the provision of communication services are based on IMS mechanisms and enablers. It makes user transparent against 3GPP CS, VoIP capable and non-VoIP capable PS and non-3GPP PS access networks; Voice Call Continuity (VCC) ensures simultaneous activation of CS and PS radio channels to enable service continuity between CS and PS systems.

[21] Source: `http://www.3gamericas.org.`

3.17 Summary

In this chapter, we gave the basics of IP networking in the context of All-IP networking. IP protocol provides specifications for routing and addressing. A node with an IP address is reachable from other nodes with IP address. Routers between these two peer nodes buffer and forward the packets. We introduce the routing protocols; RIP, OSPF, and BGP. RIP and OSPF are internal routing protocols where routing is within a network maintained by a single operator, whereas BGP is external routing protocol, which defines the routing between internal networks.

Next generation networks will proliferate the number of IP nodes and sessions will require differentiated routing with respect to type of flow. Currently, in IP version 4, IP address consists of 32 bits and IP version 6 will use 128 bits to define an IP address, which can be considered as limitless address space. Also, several QoS mechanisms are introduced to differentiate the flow during routing: DiffServ, IntServ, RSVP, MPLS. The other components required to deploy a network are IP security, IP tunneling, PPP, RADIUS & DIAMETER.

WiMAX and 4G will introduce the IP based connectivity. IP address of a mobile user needs to be same regardless of the attachment point. Mobile IP mechanism maintains the IP address of a user and facilitates directing the packets of a mobile station to appropriate attachment point during handover. With Mobile IP, a node in the network always sees the mobile user with the same IP address.

The Session Initiated Protocol (SIP) and IP Multimedia Subsystem (IMS) will play key roles in the All-IP Networking. SIP brings the performance of a circuit switched system into packet switched network, whereas IMS provides hosted multimedia services for mobile users.

References

1. 3GPP TS 35.201, “Technical Specification Group Services and System Aspects; 3G Security; Specification of the 3GPP Confidentiality and Integrity Algorithms,” `http://www.3gpp.org`.
2. 3GPP TS 35.206, “3G Security Specification of the 3GPP Confidentiality and Integrity Algorithms,” `http://www.3gpp.org`.
3. 3GPP TS 23.228, “IP Multimedia Subsystem (IMS) Stage 2,” `http://www.3gpp.org`.
4. 3GPP TS 31.102, “Identity Module (USIM) application,” `http://www.3gpp.org`.
5. 3GPP2 S.P0086-B, “IMS Security Framework,” `http://www.3gpp2.org`.
6. Rekhter, Y., Li, T., “A Border Gateway Protocol 4 (BGP-4),” RFC1771.
7. Poikselka, M., *The IMS: IP Multimedia Concepts and Services,* 2nd edition, Wiley, 2006.
8. Borman, C., et al., “Robust header compression (ROHC): Framework and four profiles: RTP, UDP, ESP, uncompressed.” IETF RFC3095, July, 2001.
9. Mathis, M., et al., “TCP selective acknowledgement options,” IETF RFC208, 1996.
10. Simpson, W., “The point-to-point protocol (PPP),” IETF RFC1661, July 1994.
11. Muratore, F., *UMTS: Mobile Communications for Future,* J Wiley, 2000.
12. Perkins, C., “Mobile IP”, *IEEE Communications Magazine,* May 1997, vol. 35, no. 5, pp. 84–99.

13. Chen, W.-T., Huang, L.-C., "RSVP Mobility Support: A Signalling Protocol for Integrated Services Internet with Mobile Hosts," INFOCOM vol. 3, pp. 1283–1292, 2000.
14. Montenegro, G., "Reverse Tunneling for Mobile IP," RFC3024, January 2001.
15. Gosh, D., Sarangan, V., Acharya, R., "Quality of Service Routing in IP Networks," *IEEE Trans. on Multimedia*, vol. 3, no. 2, pp. 200–208, June 2001.
16. Taylor, D., Herkersdorf, A., Doring, A., Dittman, G., "Robust Header Compression (ROHC) in Next Generation Network Processors" to appear in *IEEE/ACM Trans. on Networking.*
17. Bannister, J., Mather, P., Coope, S., *Convergence Technologies for 3G Networks: IP, UMTS, EGRPS and ATM,* Wiley, 2004.
18. Aissi, S., Dabbous, N., Prasan, A. R., *Security for Mobile Networks and Platforms,* Artech House, 2006.
19. Harkins, D. C., "The Internet Key Exchange (IKE)," RFC2409, November 1998.
20. Black, U., *Internet Security Protocols: Protecting IP Traffic,* Prentice Hall, 2000.
21. Frankel, S., *Demystifying IPsec Puzzle,* Artech House, 2001.
22. Holdrege, M., Srisuresh, P., "Protocol Complications with IP Network Address Translation," RFC3027, January 2001.
23. Smith, R. E., *Authentication: From Passwords to Public Keys,* Addison Wesley, 2002.
24. "PPP PAP and CHAP," RFC1334, October 1992.
25. Calhun, P., et al., "DiAMETER Base Protocol," RFC3588, September 2003.
26. Aboba, B., Simon, D., "PPP EAP TLS Authentication Protocol," RFC2716, October 1999.
27. Funk, P., Blake-Wilson, S., "EAP Tunneled TLS Authentication Protocol (EAP-TTLS)," Draft, draft-ietf-pppext-eap-ttls-03, IETF, August 2003.
28. *IEEE Standard 802.16-2004, Part 16: Air interface for fixed broadband wireless access systems,* June 2004.
29. *IEEE Standard 802.16e-2005, Part 16: Air interface for fixed and mobile broadband wireless access systems*, December 2005.
30. Johnston, D., Walker, J., "Overview of 802.16 Security," *IEEE Security and Privacy,* pp. 40–48, May/June 2004.
31. Zuleger, H., "Mobile Internet Protocol v6 (MIPv6)," `http://www.hznet.de/ipv6/mipv6-intro.pdf.`
32. Aparicio, A. C., "Analysis of the handover in a WLAN MIPv6 scenario," Global IPv6 Summit, Barcelona 2005.
33. Ergen, M., Puri, A., "MEWLANA-Mobile IP Enriched Wireless Local Area Network Architecture," *IEEE VTC,* Vancouver September, 2002.
34. Ergen, M., et al., "Position Leverage Smooth Handover Algorithm For Mobile IP," *IEEE ICN* Atlanta, August, 2002.
35. Boman, K., et al., "UMTS Security," *Electronics and Communication Engineering Journal,* vol. 14, no. 5, pp. 191–204, October 2002.
36. Tulloch, M., *Microsoft Encyclopedia of Security,* Microsoft Press, 2003.
37. "Digging Deeper into Deep Packet Inspecton," Allot Communications, 2007.
38. Arkko, J., et al., "Using IPSec to protect mobile IPv6 signalling between mobile nodes and home agents," IETF RFC3776, June 2004.
39. Balakrishna, H., et al. "A comparison of mechanisms for improving TCP performance over wireless links," *Proceedings of ACM/IEEE Mobicom,* pp. 77–89, September 1997.
40. Jacobson, V., "Compressing TCP/IP headers for low-speed serial links," IETF RFC1144, February 1990.
41. Johnson, D., Perkins, C., Arkko, J., "Mobility Support for IPv6," IETF RFC3775, June 2004.

Chapter 4
Principles of OFDM

4.1 Introduction

Orthogonal Frequency Division Multiplexing (OFDM) is a multicarrier transport technology for high data rate communication system. The OFDM concept is based on spreading the high speed data to be transmitted over a large number of low rate carriers. The carriers are orthogonal to each other and frequency spacing between them are created by using the Fast Fourier transform (FFT).[1]

OFDM originates from Frequency Division Multiplexing (FDM), in which more than one low rate signal is carried over separate carrier frequencies (Table 4.1). In FDM, separation of signal at the receiver is achieved by placing the channels sufficiently far apart so that the signal spectra does not overlap. Of course, the resulting spectral efficiency is very low as compared with OFDM, where a comparison is depicted in Fig. 4.1. Also, Fig. 4.2 shows an analogy of OFDM against single carrier and FDM in terms of spectral efficiency.

FDM is first utilized to carry high-rate signals by converting the serial high-rate signal into parallel low bit streams. Such a parallel transmission scheme when compared with high-rate single carrier scheme is costly to build. On the other hand, high-rate single carrier scheme is more susceptible to intersymbol interference (ISI). This is due to the short duration of the signal and higher distortion by its wider frequency band as compared with the long duration signal and narrow bandwidth subchannels in the parallel system.

The major contribution to the FDM complexity problem was the application of the FFT to the modulation and demodulation processes. Fortunately, this occurred at the same time digital signal processing techniques were being introduced into the design of modems.

[1] Originally Weinstein and Ebert introduced discrete Fourier transform (DFT) to create orthogonal waveforms. FFT is an efficient implementation of DFT and become defacto method with advanced very-large-scale integration (VLSI) technology. FFT reduces the number of multiplication from N^2 to $N/2\log N$ for radix-2 and $3N/8\log_2(N-2)$ for radix-4 schemes, where N is the number of orthogonal channels. Typically complexity of additions that is necessary is not significant compared with multiplication complexity.

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_4, © Springer Science+Business Media LLC 2009

Table 4.1 OFDM history (source: Wikipedia)

1957	Kineplex, multicarrier HF modem
1966	Chang, Bell Labs: OFDM paper and US patent 3488445
1971	Weinstein and Ebert proposed use of FFT and guard interval
1985	Cimini described the use of OFDM for mobile communications
1985	Telebit Trailblazer Modem incorporates a 512-carrier Packet Ensemble Protocol
1987	Alard and Lasalle: Coded OFDM for digital broadcasting
1988	TH-CSF LER, first experimental Digital TV link in OFDM, Paris area
1989	OFDM international patent application PCT/FR 89/00546, filed in the name of THOMSON-CSF, Fouche, de Couasnon, Travert, Monnier and others
1990	TH-CSF LER, first OFDM equipment field test, 34 Mbps in a 8-MHz channel, experiments in Paris area
1990	TH-CSF LER, first OFDM test bed comparison with VSB in Princeton, USA
1992	TH-CSF LER, second generation equipment field test, 70 Mbit/s in a 8-MHz channel, twin polarizations. Wuppertal, Germany
1992	TH-CSF LER, second generation field test and test bed with BBC, near London, UK
1993	TH-CSF show in Montreux SW, 4 TV channel and one HDTV channel in a single 8-MHz channel
1993	Morris: Experimental 150 Mbit/s OFDM wireless LAN
1994	US patent 5282222, method and apparatus for multiple access between transceivers in wireless communications using OFDM spread spectrum
1995	ETSI Digital Audio Broadcasting standard EUreka: First OFDM-based standard
1997	ETSI DVB-T standard
1998	Magic WAND project demonstrates OFDM modems for wireless LAN
1999	IEEE 802.11a wireless LAN standard (Wi-Fi)
2000	Proprietary fixed wireless access (V-OFDM, Flash-OFDM, etc.)
2002	IEEE 802.11g standard for wireless LAN
2004	IEEE 802.16-2004 standard for wireless MAN (WiMAX)
2004	ETSI DVB-H standard
2004	Candidate for IEEE 802.15.3a standard for wireless PAN (MB-OFDM)
2004	Candidate for IEEE 802.11n standard for next-generation wireless LAN
2005	OFDMA is candidate for the 3GPP Long Term Evolution (LTE) air interface E-UTRA downlink.
2007	The first complete LTE air interface implementation was demonstrated, including OFDM-MIMO, SC-FDMA and multi-user MIMO uplink

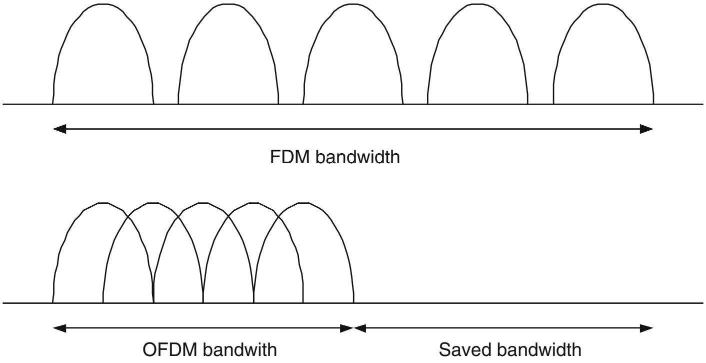


Fig. 4.1 Comparison of OFDM and FDM

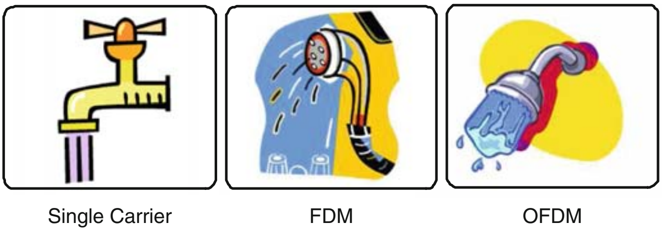


Fig. 4.2 Comparison of OFDM over FDM and single-carrier systems. OFDM and FDM are resilient to interference, since flow of water can be easily stopped in single-carrier systems. OFDM is more spectral efficient than FDM, since it utilizes the surface effectively with adjacent tiny streams

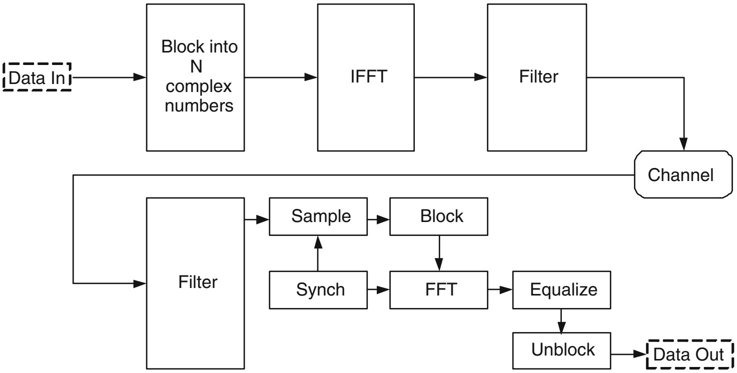


Fig. 4.3 A very basic OFDM system

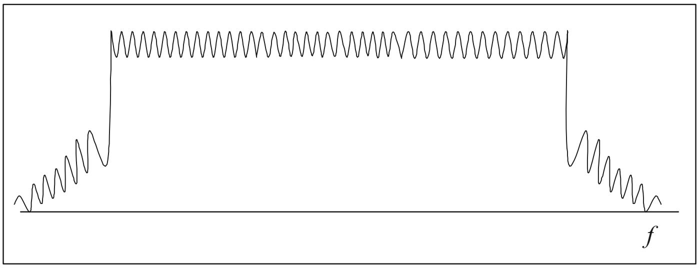


Fig. 4.4 Spectrum of OFDM signal

The technique involved assembling the input information into blocks of N complex numbers, one for each sub-channel as seen in Fig. 4.3. An inverse FFT is performed on each block, and the resultant transmitted serially. At the receiver, the information is recovered by performing an FFT on the received block of signal samples.

The spectrum of the signal on the line is identical to that of N separate QAM signals as seen in Fig. 4.4, where N frequencies separated by the signalling rate.

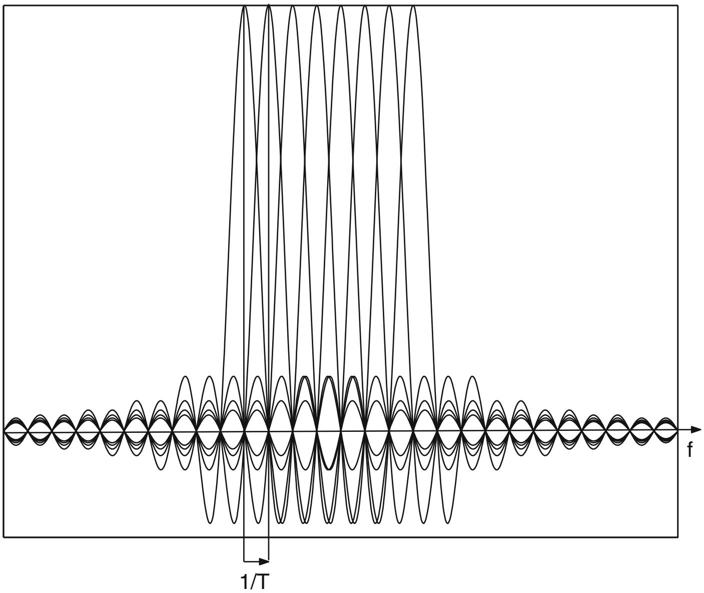


Fig. 4.5 OFDM spectrum for each QAM signal

Each QAM signal carries one of the original input complex numbers. The spectrum of each QAM signal is of the form $\frac{\sin(kf)}{f}$, with nulls at the center of the other subcarriers as seen in Fig. 4.5. This ensures orthogonality of subcarriers.

However, orthogonality is threatened by intersymbol interference (ISI), which is caused by leakage of symbols into another due to multipath interference. To combat for ISI, a guard time is introduced before the OFDM symbol. Guard time is selected longer than impulse response or multipath delay so as not to cause interference of multipath components of one symbol with the next symbol.

Orthogonality is also threatened by intercarrier interference (ICI), which is crosstalk between subcarriers, since now the multipath component of one subcarrier can disturb the another one. ICI in OFDM is prevented by cyclically extending the guard interval as seen in Fig. 4.6 to ensure integer number of cycles in the symbol time as long as the delay is smaller than the guard time.

Another issue is how to transmit the sequence of complex numbers from the output of the inverse FFT over the channel. The process is straightforward if the signal is to be further modulated by a modulator with I and Q inputs as in Fig. 4.7.

Otherwise, it is necessary to transmit real quantities. This can be accomplished by first appending the complex conjugate to the original input block. A $2N$-point inverse FFT now yields $2N$ real numbers to be transmitted per block, which is equivalent to N complex numbers.

OFDM increases the robustness against frequency selective fading or narrowband interference due to narrowband flat fading subchannels. As compared with single carrier system, a single fade or interferer can cause the entire link to fail, but in

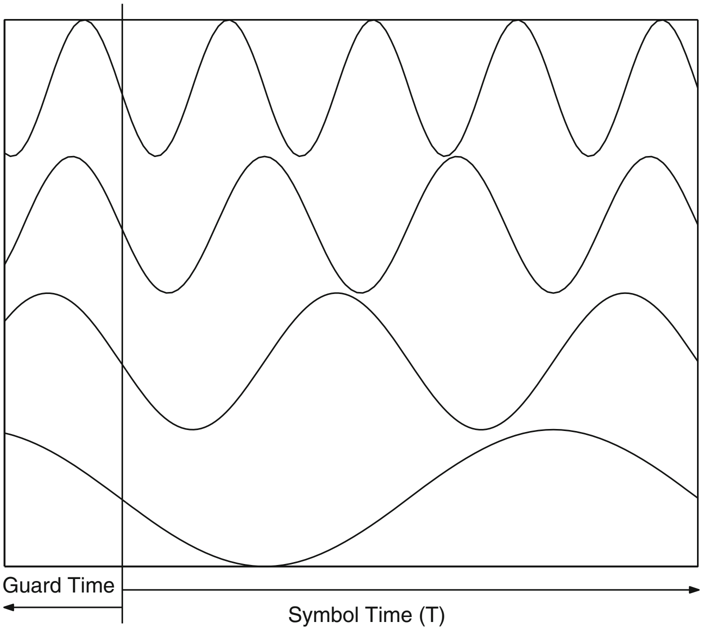


Fig. 4.6 OFDM with cyclic shift

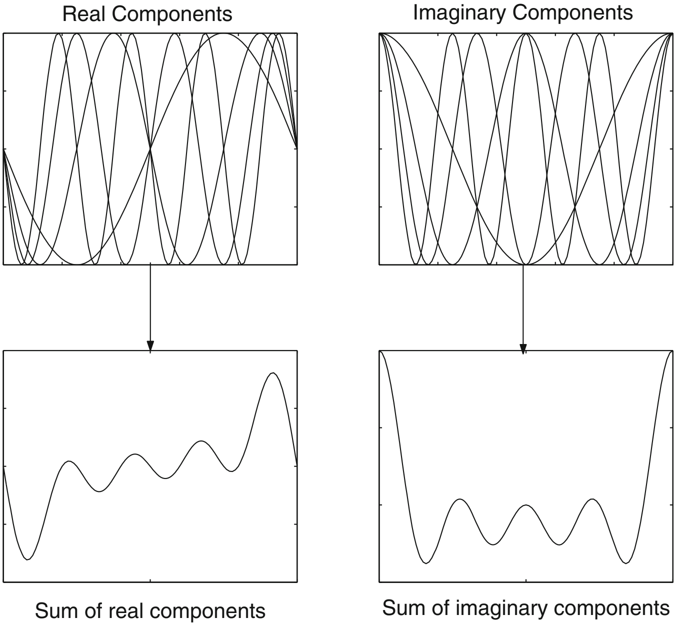


Fig. 4.7 Real and Imaginary components of an OFDM symbol: The superposition of several harmonics modulated by data symbols

OFDM, since there are several subcarriers, only small percentage of the subcarriers is affected. Error correction coding is used to correct the erroneous subcarriers.

OFDM on the other hand suffers from noise such as amplitude with a very large dynamic range; therefore, it requires RF power amplifiers with a high peak to average ratio. It is also more sensitive to carrier frequency offset than single carrier systems are due to leakage of the FFT.

OFDM has been particularly successful in numerous wireless applications, where its superior performance in multipath environments is desirable. Wireless receivers detect signals distorted by time and frequency selective fading. OFDM in conjunction with proper coding and interleaving is a powerful technique for combating the wireless channel impairments that a typical OFDM wireless system might face.

4.2 A simple OFDM system

Let us consider Fig. 4.8 as a simple OFDM system to understand the mechanics behind it. The incoming data is converted from serial to parallel and grouped into bits each to form a complex number x after PSK or QAM modulation in order to be transmitted over N low-rate data streams. Each low-rate data stream is associated with a subcarrier of the form

$$\phi_k(t) = e^{j2\pi f_k t}, \tag{4.1}$$

where f_k is the frequency of the kth subcarrier. Consequently, one baseband OFDM symbol with N subcarrier is

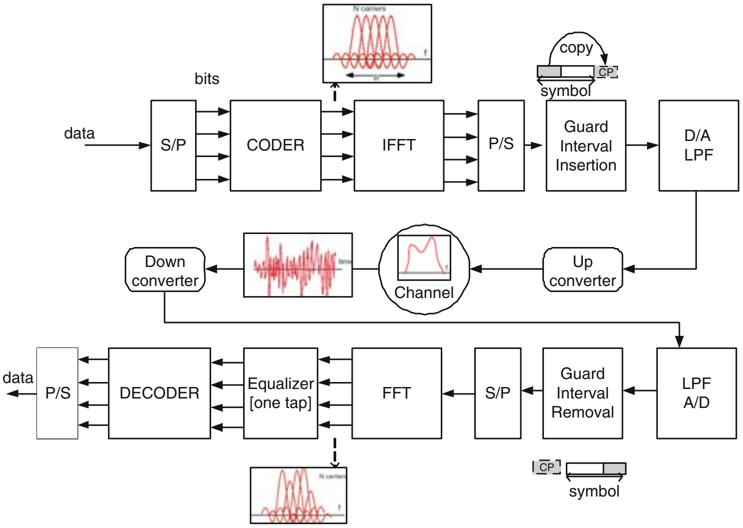


Fig. 4.8 Simplified OFDM system

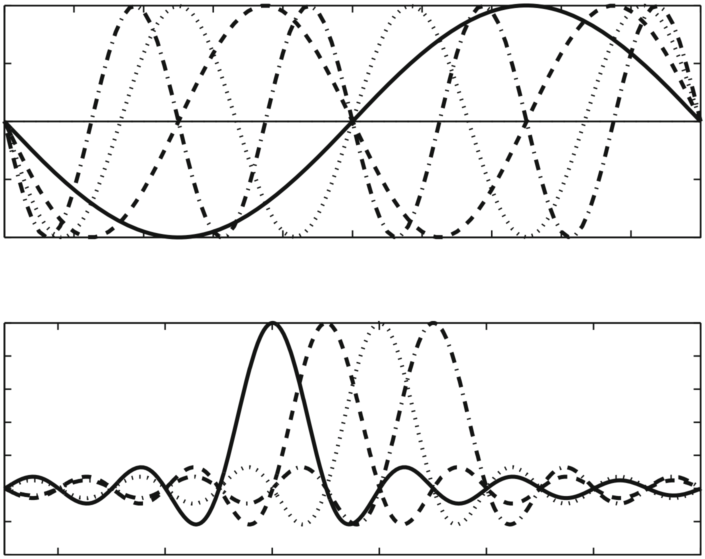


Fig. 4.9 An example of four subcarriers in time and frequency with same modulation

$$s(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_k \phi_k(t), \quad 0 < t < T, \tag{4.2}$$

where x_k is the kth complex data symbol and T is the length of the OFDM symbol.

Now, let us look at the constructed OFDM symbol in detail to analyze the orthogonality of subcarriers. Consider four subcarriers ($N = 4$), and first assume same modulation for each subcarrier. Figure 4.9 shows those four subcarriers in time and frequency. Also consider Fig. 4.10 for an example to highlight the effect of different modulation in each subcarrier.

Orthogonality of OFDM subcarriers in frequency domain is with Dirac pulses convolved with $\text{sinc}(\pi f T)$. Since in time domain a subcarrier ϕ_k is multiplied with a rectangle(T), which is in frequency domain a convolution between $\delta(f - f_k)$ and $\text{sinc}(\pi f T)$. This is basically $1/T$ shifted version of $\text{sinc}(f)$ for each f_k and $\text{sinc}(\pi f T)$ has zeros for all frequencies that are integer multiple of $1/T$.

Notice that the orthogonality of OFDM subcarriers can be demonstrated in time domain as well. Within a symbol time (T), there are integer number of cycles in the symbol interval, and the number of cycles between adjacent subcarriers differs exactly by one (see first subfigures in Figs. 4.9 and 4.10). In the receiver when it is demodulated, down converted, and integrated with a frequency j/T, then the x_j is received since any other subcarrier when it is down converted with a frequency $(i - j)/T$ produces zero after integration since $(i - j)/T$ produces integer number of cycles within the integration interval (T).

When signal is transmitted over a channel, channel dispersion destroys the orthogonality between subcarries and causes ICI, and delay spread causes ISI between

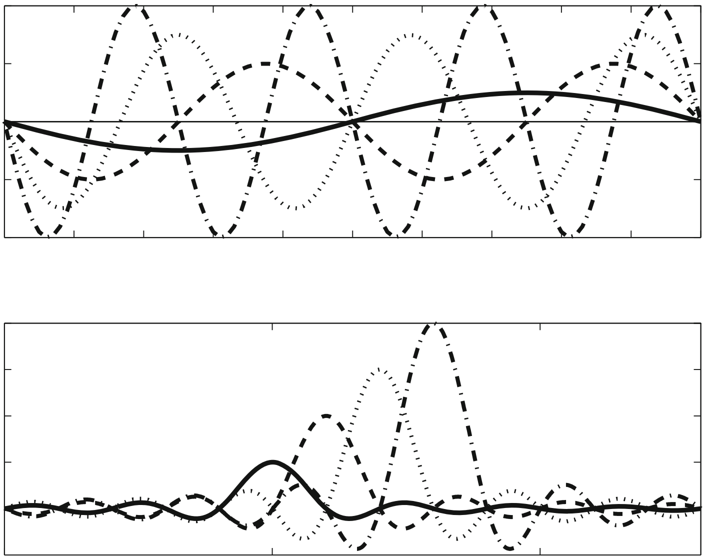
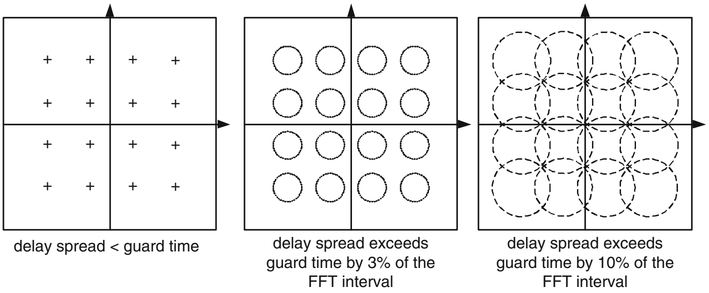


Fig. 4.10 An example of four subcarriers in time and frequency with different modulation: Modulation level increases with the increasing number of subcarriers



delay spread < guard time
delay spread exceeds guard time by 3% of the FFT interval
delay spread exceeds guard time by 10% of the FFT interval

Fig. 4.11 16QAM constellation

successive OFDM symbols. As we mentioned before, cyclic prefix (CP) is used to preserve the orthogonality and avoid ISI. We will see that this makes equalization in the receiver very simple. If multipath exceeds the CP, then constellation points in the modulation is distorted. As can be seen from Fig. 4.11, when multipath delay exceeds the CP, the subcarriers are not guaranteed to be orthogonal anymore, since modulation points may fall into anywhere in the respective contour. As delay spread gets more severe, the radius of the contour enlarges and crosses the other contours. Hence, this causes error.

The CP is utilized in the guard period between successive blocks and constructed by the cyclic extension of the OFDM symbol over a period τ:

$$s(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} x_k \phi_k(t), \quad -\tau < t < NT. \tag{4.3}$$

The required criteria is that τ is chosen bigger than channel length τ_h so as not to experience an ISI. The CP requires more transmit energy and reduces the bit rate to $(Nb/NT+\tau)$, where b is the bits that a subcarrier can transmit.

The CP converts a discrete time linear convolution into a discrete time circular convolution. Thus, transmitted data can be modeled as a circular convolution between the channel impulse response and the transmitted data block, which in the frequency domain is a pointwise multiplication of DFT samples. Then received signal becomes

$$y(t) = s(t) * h(t) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} H_k x_k \phi_k(t), \quad 0 < t < NT, \tag{4.4}$$

where

$$H_k = \int_0^{\tau_h} h(t) e^{j2\pi f_k t} \mathrm{d}t. \tag{4.5}$$

Hence, kth subcarrier now has a channel component H_k, which is the fourier transform of $h(t)$ at the frequency f_k.

The OFDM symbol is sampled ($t = nT$ and $f_k = k/NT$) in the receiver and demodulated with an FFT. Consequently, the received data has the following form

$$y_k = H_k x_k, \quad k = 0, \ldots, N-1. \tag{4.6}$$

The received actual data can be retrieved with N parallel one-tap equalizers. One-tap equalizer simply uses the estimated channel ($\hat{H}_k$) components and use it to retrieve estimated $\hat{x}_k$ as follows

$$\hat{x}_k = \frac{y_k}{\hat{H}_k} = \frac{H_k}{\hat{H}_k} x_k. \tag{4.7}$$

Also note that the spectrum of OFDM decays slowly. This causes spectrum leakage to neighboring bands. Pulse shaping is used to change the spectral shape by either commonly used raised cosine time window or passing through a filter.

An OFDM system design considers setting the guard interval (τ) as well as the symbol time (T) and FFT size with respect to desired bit rate B and given tolerable delay spread. The guard interval is selected according to delay spread, and typically it is 2–4 times the root-mean-squared delay spread with respect to chosen coding and modulation.

Symbol time is set with respect to guard time and it is desirable to select much larger than the guard time since the loss in SNR in the guard time is compensated. Symbol time as we know determines the subcarrier spacing $(f_b = 1/T)$. Number of subcarriers N is found with respect to desired bit rate, since total number of bits (b_T) to carry in one symbol is found with $B/(T+\tau)$ and selected coding and modulation determines the number of bits (b) in one subcarrier. Hence, the number of subcarriers is $N = b_T/b$. For instance, b is two for 16QAM with rate $1/2$. The required bandwidth (W) is then $N * f_b$. Alternatively, this method is reversed to find out the symbol time starting from the given bandwidth.

This section described how the basic OFDM transceiver is formed. However, there are more components to build a complete OFDM transceiver. Figure 4.12

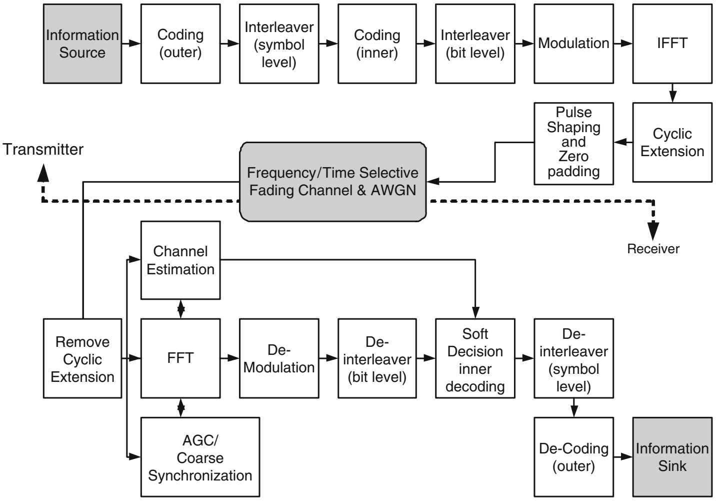

Information Source
Coding (outer)
Interleaver (symbol level)
Coding (inner)
Interleaver (bit level)
Modulation
IFFT
Cyclic Extension
Pulse Shaping and Zero padding
Transmitter
Frequency/Time Selective Fading Channel & AWGN
Receiver
Channel Estimation
Remove Cyclic Extension
FFT
De-Modulation
De-interleaver (bit level)
Soft Decision inner decoding
De-interleaver (symbol level)
AGC/ Coarse Synchronization
De-Coding (outer)
Information Sink

Fig. 4.12 A typical wireless OFDM architecture

shows the block diagram of an OFDM transceiver, where the upper path is the transmitter and the lower path corresponds to the receiver. In the remainder of this chapter, we will consider this OFDM transceiver and describe the components in detail.

4.3 Coding

In the previous section, we gave a simple uncoded OFDM system and highlighted the key features. In a multipath environment, all subcarriers will experience different fading environment and all will arrive with different amplitudes. Some of them will experience a deep fade, which will cause error during detection, and the average probability of error will be the same as that for a flat fading single-carrier system with the same average geometric mean of the subcarriers' SNRs. These errors may dominate the bit error rate. To avoid this domination, error correction coding is utilized. Coding[2] used in OFDM systems is to correct the certain number of errors in order to enable high rates in the presence of fading and interference from other wireless channels.

[2] "Why use error coding? Error coding may be selected to improve data reliability, reduce system costs, or increase the range. For instance, 3 dB coding gain can

- increase throughput 2-fold or
- increase range by 40% or
- reduce bandwidth by 50% or
- reduce antenna size by 30% or
- reduce transmitter power by half."

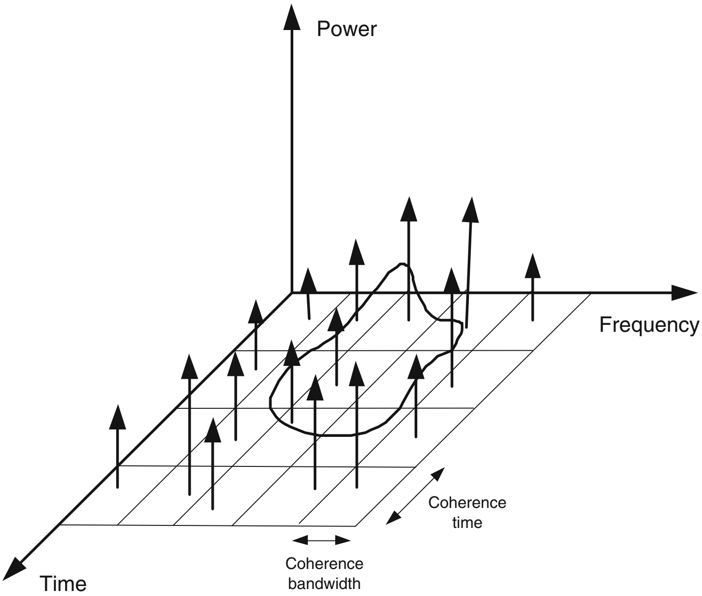


Fig. 4.13 Two-dimensional coding for OFDM with respect to channel impulse response

This section starts with an introduction of block and convolutional coding. Then, we introduce concatenated coding as a way to combine different coding schemes to further reduce the error rate. A combination of block coding and convolutional coding along with proper time and frequency interleaving as seen in Fig. 4.13 constitutes such a concatenated coding strategy to achieve better frequency diversity.[3]

In the past several years, convolutional coding with Viterbi decoding has begun to be supplemented in the geostationary satellite communication arena with Reed-Solomon coding. The two coding techniques are usually implemented serially as concatenated block and convolutional coding. Typically, the information to be transmitted is first encoded with the Reed-Solomon code, then with the convolutional code. On the receiving end, Viterbi decoding is performed first, followed by Reed-Solomon decoding. This is the technique that is used in most if not all of the direct-broadcast satellite (DBS) systems, and in cellular communication as well.

Later, we review some fundamental features of turbo coding and explain why it is a good potential candidate for OFDM systems. Turbo coding is a new parallel-concatenated convolutional coding technique. Initial hardware encoder and decoder implementations of turbo coding have already appeared on the market. This

[3] But achievable frequency diversity is limited by number of resolvable independent paths in the channel impulse response. Intuitive explanation is as follows: if the channel is one-tap, the SNR on all subcarriers is the same since channel length is small compared with the OFDM symbol that makes the subcarries correlated. If the number of resolvable taps increases, the correlation between subcarriers decreases and diversity increases. But resolvable number of taps are limited, and therefore subcarriers cannot be independent.

technique achieves substantial improvements in performance over concatenated Viterbi and Reed-Solomon coding. Turbo coding has recently been used in many communication systems successfully. Random-like structure of turbo codes have resulted in outstanding performance by providing small error rates at information rates of close to theoretical channel capacity.

Finally, we conclude the section with Trellis coding, which is a coding technique that also leverages modulation and low-density parity check (LDPC) coding, which is becoming popular in cellular communication and included as a supported coding scheme in several standards.

4.3.1 Block Coding

Block coding is a way of mapping k symbols to n symbols with $n > k$. We call the block of k symbols a messageword, and the block of n symbols a codeword. The process of mapping k message symbols to n code symbols is called *encoding*, and the reverse process is called *decoding*.

This mapping can be systematic (Fig. 4.14), wherein the block of n codeword symbols consists of the k messageword symbols plus $(n-k)$ added redundant symbols, or nonsystematic, where the messageword symbols cannot be directly recovered from the codeword without decoding. Generally, any linear block code can be represented as an "equivalent" systematic code. A block code is called "linear" if the sum of two codewords is always a valid codeword, and a scalar multiple of any codeword is also a valid codeword.

Block codes, in particular the Reed-Solomon class, are used to combat for bursty errors. Burst errors occur in single-carrier systems when the impulsive noise has duration greater than a symbol period and coherence time of the channel is longer than symbol period.

In OFDM system, an impulsive noise with a wide-frequency content causes burst error to affect several adjacent subcarriers if coherence bandwidth of the channel is wider than the subcarrier spacing.

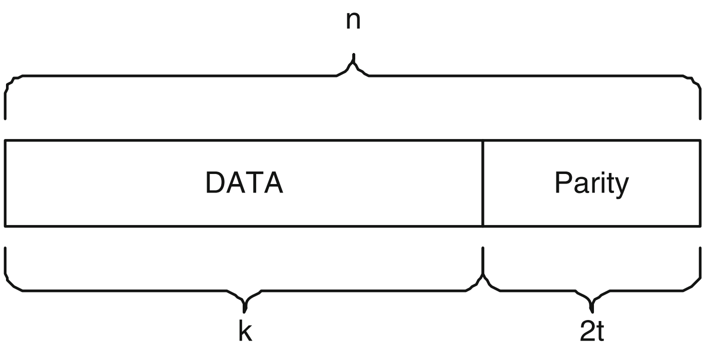


Fig. 4.14 Construction of a systematic block code

4.3.1.1 Interleaving

The performance of codes in the presence of bursts can be improved by the process of interleaving. To separate the correlated components, the code is made to operate on symbols with sufficient spacing so that the errors are more independent. At the receiver, the symbols are de-interleaved before decoding. The decoder therefore operates on symbols spaced some symbol periods apart as transmitted. Figure 4.15 shows a block interleaving method for transmitter and receiver. This interleaver writes by row and reads by column. Pseudo-random interleaver on the other hand reads the information bits to the encoder in a random but fixed order.

Interleaving can be in time, frequency, or both. For instance, for time interleaving, channel coherence time is on the order of 10–100 s of symbols, which makes the channel highly correlated across adjacent symbols. Interleaving makes sure that symbols especially adjacent symbols experience nearly independent fading gains.

4.3.1.2 Cyclic Redundancy Check

Cyclic redundancy check (CRC) is a simple form of block coding for error detection. Fixed number of check bits follows messageword. If the receiver detects an error, a retransmission is requested. Figure 4.16 shows a typical encoder implementation of CRC with an n-bit feedback shift register whose connection pattern is a primitive

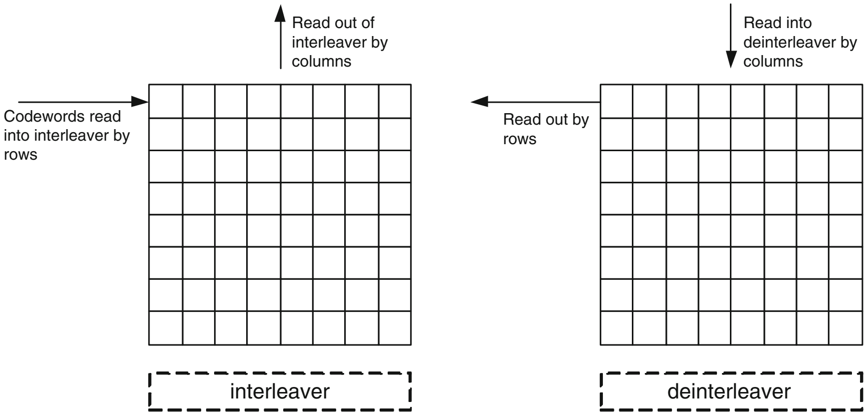

Read out of interleaver by columns
Codewords read into interleaver by rows
interleaver
Read into deinterleaver by columns
Read out by rows
deinterleaver

Fig. 4.15 Implementation of block interleaving

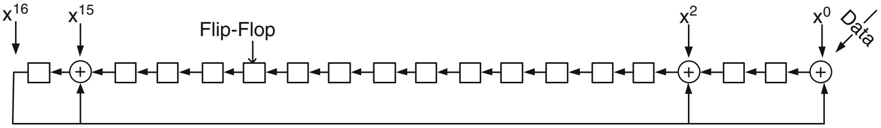


Fig. 4.16 CRC-16 implementation: $P(x)= x^{16}+x^{15}+x^{2}+1$

polynomial. At the encoder, shift register starts with a predetermined pattern, and input data is fed both to the channel and to the feedback shift register. At the receiver, received data and output is concomitantly fed back to the register. A CRC of length n can detect any error pattern of length less than n with probability $1-2^{-n}$.

Let's now look at how coding is used to correct errors. In brief, the idea behind error correction coding is to start with a "message" (i.e., the thing you want to encode) of length k, and convert it to a "codeword" of longer length n, in such a way that the additional information in the coded form allows one to recover the original message if parts of it are corrupted. To see how this works, we will need some additional definitions:

4.3.1.3 Hamming weight

The *Hamming weight* of a codeword is a metric that indicates the number of nonzero symbols in codeword. For example, if the codeword is 100010001, its weight would be 3. If the codeword is the nonbinary 23012001, then the weight would be 5. It works the same way regardless off the base field.

4.3.1.4 Hamming Distance

The *Hamming distance* is a comparison metric between two codewords by the number of places where the codewords differ. So, for example, given the two binary codewords 100111 and 110000, the Hamming distance between them would be 4.

4.3.1.5 Minimum distance of a code

The *minimum distance of a code* is another metric that typically gives a characteristic of the code by measuring the minimum distance between all the codewords in the code. This is achieved by taking the distance between each codeword and every other codeword in the code, and the minimum gives the minimum distance of the code. For linear codes, minimum distance equals the lowest hamming weight in the code. Error correction capability of the code is highly correlated with the minimum distance of a code. Consider a simple binary repetition code of length 4, where there are two codewords (1111) and (0000). The minimum distance is 4, since minimum hamming weight is 4. Suppose we send (1111):

Transmitted	Received	Hamming distance to (1111)	Decision
1111	0111	1	1111
1111	1010	2	Fail
1111	0001	3	0000

If the received bits have equal hamming distance to two code words, then decoder cannot decode it and fails. Otherwise, decoder decides for the codeword that has the smallest distance away from the received word. This type of decoding can be generalized to much larger and more sophisticated codes, and for a minimum distance d we can correct up to distance floor$((d-1)/2)$.

4.3.2 Reed-Solomon Coding

Reed-Solomon (RS) codes are a very popular block coding technique of today as compared with evolving capacity-approaching codes. RS is easy to decode and suits best for high-rate systems with small data packets.

4.3.2.1 Cyclic Codes

RS codes are cyclic codes in addition to being linear block codes. A cyclic code preserves the property of being cyclic in the sense that when shifted the result is still a codeword. For instance, if codeword (0111) is shifted one digit, the result (1011) is also a codeword. Consequently, all circular shifts of any codeword in the code are also codewords in the code. In polynomial terms, multiplication of cyclic codeword code(x) with x results in circular shift of the codeword x.code$(x)\,\text{mod}(x^n-1)$ and the following manipulations are possible:

$$a_{n-1}x_{n-1}c(x)+a_{n-2}x_{n-2}c(x)+\cdots+a_0c(x), \tag{4.8}$$

since the sum of two codewords is always a codeword in linear codes, and multiplication of a codeword by a scalar always results in a codeword. Notice that all operations must be done using Galois field (GF) arithmetic.[4] For example, a binary code uses two values, the binary numbers $\{0,1\}$, as symbols. In general, though, a code uses q-ary symbols. Q-ary symbols use symbols taken from an alphabet A of q possible values. So, for example, 5-ary symbols would be symbols chosen from a set of five elements, such as $\{0, 1, 2, 3, 4\}$. Practical RS codes use $q=256$, since it can be represented using 8-bit symbols per codeword.

There are two common definitions of Reed-Solomon codes: as polynomial codes over finite fields and as cyclic codes of length $q-1$ over $GF(q)$.

[4] "A field is a set of elements that can be added, subtracted, multiplied, and divided, with the important stipulation that the result of any of those operations is always still an element of the field. Additionally, we require that additive and multiplicative inverses and identity elements exist for each (non-zero) element of the field, and that the field elements obey the familiar commutative, associative, and distributive properties. The real numbers are a familiar example of a field: we can add, subtract, multiply and divide any two (non zero) real numbers, and the result is always another real number. Multiplicative and additive inverses can be found for any real number; the multiplicative identity element is '1', the additive identity element is '0'. The real numbers are an infinite field. We can construct a field with a finite number of elements, if we follow certain rules for constructing such fields."

4.3.2.2 Polynomial Codes over Certain Finite Fields

The idea[5] is very simple, there is a message $m(x)$

$$m(x) = m_{k-1}x^{k-1} + m_{k-2}x^{k-2} + \cdots + m_1 x + m_0 \tag{4.9}$$

in the form of a polynomial whose coefficients (m_i) are taken from finite field $GF(q)$. Codeword is found by evaluating the $m(x)$ at n distinct elements of the finite field:

$$(c_0, c_1, c_2, ..., c_{n-1}) = m(a_0), m(a_1), m(a_2), ..., m(a_{n-1}), \tag{4.10}$$

where n distinct elements of the field are $a_0, a_1, ..., a_{n-1}$. A generalization of the above construction leads to the definition of generalized Reed-Solomon (GRS) codes:

$$\begin{aligned}&(c_0, c_1, c_2, ..., c_{n-1}) =\\ &v_0 m(a_0), v_1 m(a_1), v_2 m(a_2), ..., v_{n-1} m(a_{n-1}),\end{aligned} \tag{4.11}$$

where $v_0, v_1, ..., v_{n-1}$ be n nonzero (but not necessarily distinct) elements of $GF(q)$.

If the message has k symbols, and the length of the code is $n = q - 1$, then the code consists of n equations in k unknowns, which is overspecified when $n > k$, hence the correct coefficients can be recovered even if some of them are corrupted.

4.3.2.3 Generator Polynomial Approach

Given the message $m(x)$ in the form of a polynomial, as outlined earlier, whose k coefficients are taken from the finite field with q elements, we can construct RS codewords with $c(x) = m(x)g(x)$ (or the equivalent systematic construction). All we need to do is specify the generator polynomial of the code.

The general form of the generator polynomial of a RS code is defined in such a way as to have its roots $2t$ consecutive powers of a primitive element. Thus we can write,

$$g(x) = (x - ab)(x - ab + 1)(x - ab + 2)...(x - ab + 2t - 1). \tag{4.12}$$

For convenience, the constant b is often chosen to be 0 or 1. Given the generator polynomial, RS codewords can now be constructed as $c(x) = m(x)g(x)$, where $g(x) = (x - a^i)(x - a^{i+1})(x - a^{i+2t-1})$ and $m(x)$ are the information element. This method is often used in practice, since polynomial multiplication is relatively easy to implement in hardware. Therefore, an RS code with $2t$ check symbols can correct up to floor$((2t + 1 - 1)/2) = t$ errors or $2t$ erasures. An erasure occurs when the position of an error symbol is known.

RS codes are the best minimum distance obtainable codes. Recall that minimum distance is the most important property of an error correction code. Since they are

[5] "Original definition used by Irving S. Reed and Gustave Solomon in their paper "Polynomial codes over certain finite fields" published in the *Journal of the Society for Industrial and Applied Mathematics* in 1960."

linear, cyclic, and their generator polynomial has well-defined roots, RS codes are easy to encode and relatively easy to decode. The coding gain with RS dictates that the probability of error correction in the decoded data is higher than the probability of error correction without Reed-Solomon. Hence, the probability of error in the remaining decoded data is lower.

A popular Reed Solomon code is RS(255,233) with 8-bit symbols, where $n = 255$, $k = 223$, and $s = 8$ and $2t = 32$. The decoder can correct any 16 symbol errors in the codeword. Larger value of t means that larger value of errors can be corrected, but it requires more computational power.

4.3.3 Convolutional Coding

Convolutional coding is another famous coding technique that operates on serial streams of symbols rather than blocks. A convolutional encoder is usually described by two parameters: the code rate and the constraint length. The code rate is k/n, where k is the number of bits into the convolutional encoder and n is the number of channel symbols output by the convolutional encoder in a given encoder cycle. The constraint length parameter, K, denotes the "length" of the convolutional encoder, which denotes the number of stages and the cycles an input bit retains in the convolutional encoder.

Viterbi[6] decoding or sequential decoding are used for convolutional encoding. Sequential decoding performs well with long-constraint-length convolutional codes, but it has a variable decoding time. Viterbi decoding on the other hand has a fixed decoding time. In hardware implementation it is preferable, but complexity of the algorithm increases exponentially with a function of constraint length. Viterbi also permits soft decision decoding, where the minimum Euclidean distance[7] between the received sequence and all allowed sequences, rather than Hamming Distance, is used to form decisions.

4.3.3.1 Encoder

To perform convolutional encoding, shift register[8] and module-two addition combinatorial logic is needed. The encoder shown in Fig. 4.17 encodes the $K = 3$, $(7,5)$

[6] "Viterbi decoding was developed by Andrew J. Viterbi, a founder of Qualcomm Corporation. the technique is presented in 'Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm,' published in IEEE Transactions on Information Theory, Volume IT-13, pages 260-269, in April, 1967."

[7] A straight line distance between any two points is called the *Euclidean distance*. Euclidean distance can be considered between sequences.

[8] "A shift register is a chain of flip–flops wherein the output of the n^{th} flip–flop is tied to the input of the $(n+1)^{\text{th}}$ flip–flop. At every active edge of the clock, output is considered input of the flip–flop, and thus the data are shifted over one stage."

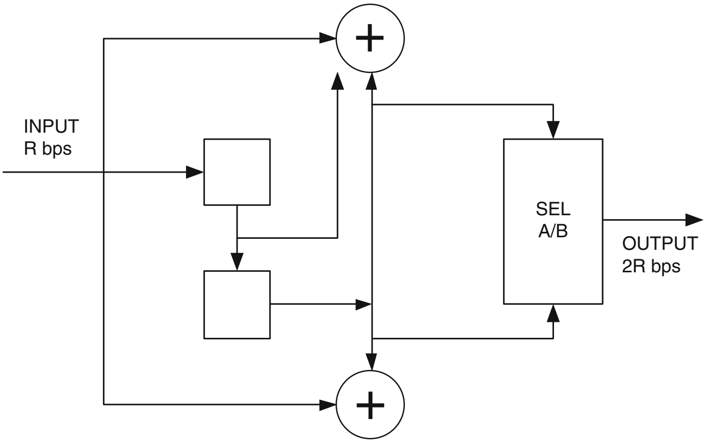


Fig. 4.17 Generation of a convolutional code: data bits are provided at a rate of R bits per second. Channel symbols are output at a rate of $2R$ symbols per second. The input bit is stable during the encoder cycle. When the input clock edge occurs, the output of the left-hand flip–flop is clocked into the right-hand flip–flop, the previous input bit is clocked into the left-hand flip–flop, and a new input bit becomes available. Then the outputs of the upper and lower modulo-two adders become stable. The output selector (SEL A/B block) cycles through two states-in the first state, it selects and outputs the output of the upper modulo-two adder; in the second state, it selects and outputs the output of the lower modulo-two adder

Current State	Next State	
	Input=0	Input=1
00	00	10
01	00	10
10	01	11
11	01	11

Current State	Output	
	Input=0	Input=1
00	00	11
01	11	00
10	10	01
11	01	10

Fig. 4.18 State diagram of a convolutional code: "These two tables are enough to describe the behavior of the example rate 1/2, $K = 3$ convolutional encoder"

convolutional code where rate is $1/2$ and $m = 2$. The octal numbers 7 and 5 represent the code generator polynomials, which when read in binary (111_2 and 101_2) correspond to the shift register connections to the upper and lower modulo-two adders, respectively.

Each input bit has an effect on three successive pairs of output symbols. That gives the convolutional code its error-correcting power. The example encoder has two bits of memory, and so there are four possible states. In the state diagram shown in Fig. 4.18, the transitions from a current state to a next state are determined by the current input bit, and each transition produces an output of n bits.

The code rate of convolutional coding can be increased by puncturing. Puncturing removes some of the bits after encoding. This gives the same code rate as if it is encoded with a higher rate code. But with puncturing, any code rate is achievable with the same decoder.

In addition, interleaving may be applied to a convolutional coding as in block coding, since Viterbi algorithm operates at optimum if the received inputs are independent. Interleaving eliminates the correlation in time and frequency and sufficiently provides independent received inputs by converting Rayleigh channel to an approximated Gaussian one. Figure 4.19 illustrates the performance of an interleaved convolutional code over a Rayleigh channel. Notice that interleaving provides a far lower error rate and steeper curve.

4.3.3.2 Viterbi Decoder

The Viterbi algorithm operates on the Trellis diagram. Figure 4.20 shows the Trellis diagram for rate 1/2 $K = 3$ convolutional encoder, for a 15-bit message. The operation consists of finding the path of states, where a state denotes the past history of the sequence over the constraint length as described earlier. This Viterbi decoder presented in the figure operates on hard decision and so considers only hamming distance. For soft decoding, the algorithm would find the transmitted sequence whose

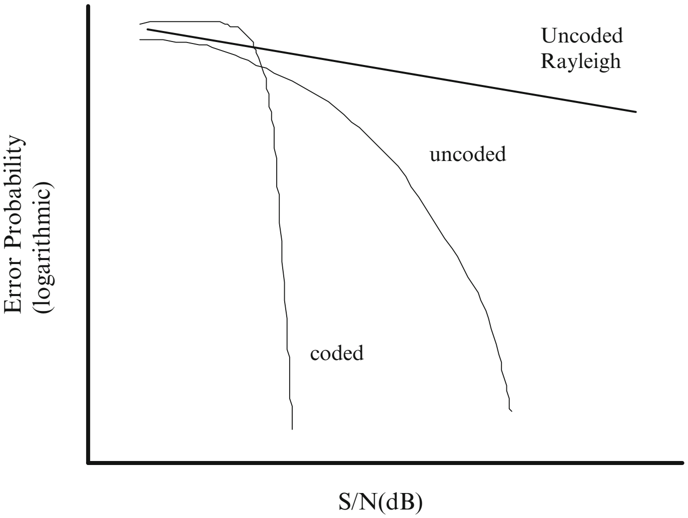


Fig. 4.19 Performance of a convolutional code over a Rayleigh fading channel

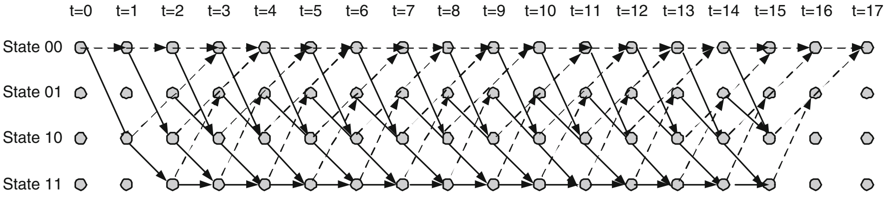


Fig. 4.20 Trellis diagram

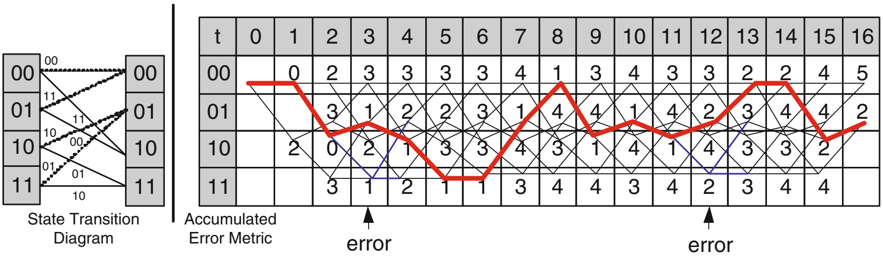


Fig. 4.21 Viterbi decoding process

Euclidean distance, or equivalently whose accumulated square distance, is closest to the received sequence.

Figure 4.20 shows the Trellis diagram with four states. Notice that encoder is flushed to the initial state with two memory flushing bits (two zeroes) appended to 15-bit message. If input is one, it is represented by solid lines and otherwise it is represented by dotted lines. Notice that arrows are in line with the state transition table presented in Fig. 4.18.

Viterbi decoder decides about the original bits according to accumulated error metric. At time instant t, the smallest accumulated error metric is selected according to the history of what states preceded the states. This method is called *traceback* method.[9]

Accumulated error metric for each branch is shown in Fig. 4.21. It also shows the state transition diagram. Notice that state transition diagram shows the restricted transition between states, and solid lines are used for input one and dashed lines are for input zero. Viterbi decoder selects the state that has the smallest accumulated error metric and iteratively performs backward.

For each branch, there is accumulated error metric and minimum is selected. If accumulated error metric is equal, then the decoder decides by looking forward. For instance, at time $t = 3$ and $t = 12$, there are errors, and the decoder can end up with the same accumulated error. Next forward step shows in both cases the correct state.

4.3.4 Concatenated Coding

Concatenated coding has emerged from the need to achieve better error correcting capabilities with less-complex structures. The concatenated coding uses two or more codes after each other or, in parallel, usually with some kind of interleaving. The constituents of the codes are decoded with their own decoders.

[9] "Traceback is not scalable with longer messages due to memory constraints and decoder delay. A *traceback* depth of K × 5 is sufficient for Viterbi decoding. Source: C. Fleming, "A Tutorial on Convolutional Coding with Viterbi Decoding," Spectrum Applications, Copyright 1999-2006."

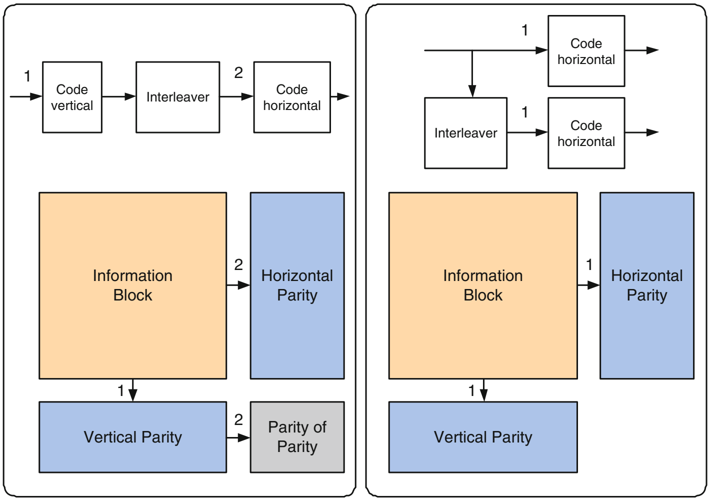


Fig. 4.22 Concatenated coding

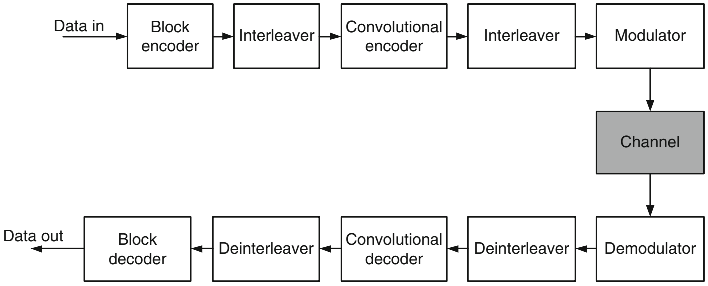


Fig. 4.23 Concatenated coding with interleaving

The idea of concatenated coding is illustrated in Fig. 4.22. Serial and parallel concatenated coding is illustrated. Notice that in serial concatenated coding, parity bits generated from the first encoding are also encoded in the second code. We do not see this parity of parity if codes are working in parallel.

Concatenated coding, which combines the block coding and convolutional coding, is illustrated in Fig. 4.23. This structure effectively combats for errors. The block coding is applied before convolutional coding and block decoding is applied after convolutional decoding. Interleaving is performed in between for superior performance with different interleaving patterns.

The inner convolutional code performs superior error correction with soft decision decoding, and if convolutional code makes an error, it causes a large burst, since Viterbi algorithm may pick a wrong sequence. In this case, we know that block coding, especially an interleaved Reed-Solomon coding, is superior correcting the bursty errors.

Concatenated coding provides means to constructing long codes, and it also confirms Shannon's channel coding theorem by stating that if the code is long enough, any error can be corrected.

4.3.5 Trellis Coding

Trellis coding[10] in simplest terms is a combination of coding and modulation. Coding and modulation is gelled together neatly to show that coded modulation schemes are capable of operation near Shannon capacity on bandlimited channels. Soft decision decoding is based on minimum Euclidean distance of sequences and is part of the demodulation procedure.

Trellis coding adds redundant constellation points rather than redundant bits or symbols. Consequently, bit rate increases but the symbol rate stays the same and it conserves bandwidth. Increasing the constellation size reduces Euclidean distance between the constellation points, but sequence coding offers a coding gain that overcomes the power disadvantage of going to higher constellation.

Figure 4.24 shows Trellis-coded modulation. The first step in designing a Trellis code is to form an expanded constellation and to partition it into subsets. The points within each subset are made far apart in Euclidean distance, and will correspond to uncoded bits. The remaining, or coded bits, determine the choice of subset. The code rate is different and each adds *m* extra bit to the symbol bit size.

Figure 4.25 shows Trellis coding for QAM modulation, which adds one extra bit and expands the constellation without increasing the signal energy. The signal energy is kept the same, since the distance between the symbols decreases. Although it sounds like a disadvantage in performance, advantage comes from the restriction on what transitions are allowed in the constellation. Those transitions are being

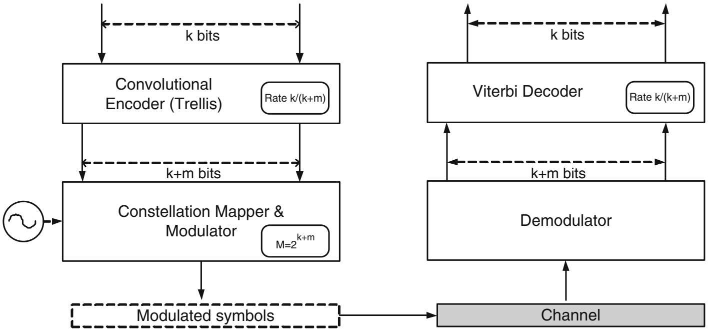


Fig. 4.24 Trellis coded modulation

[10] Invented by Gottfried Ungerboeck in 1982. This technique is used in telephone-line modems to squeeze high ratios of bits-per-second to Hertz out of 3 KHz-bandwidth analog telephone lines. The Viterbi decoding algorithm is also used in decoding Trellis-coded modulation as well.

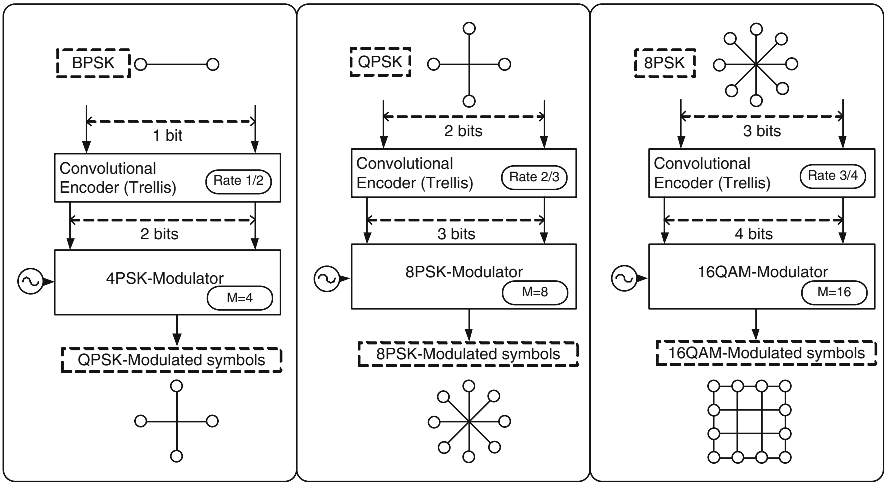


Fig. 4.25 Trellis-coded modulation: BPSK: code rate 1/2, output QPSK; QPSK code rate 2/3, output 8PSK; 8PSK, code rate = $3/4$, output 16QAM

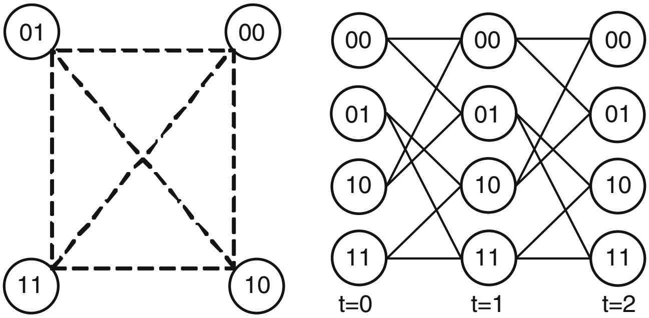


Fig. 4.26 QPSK with and without Trellis coding

determined by a simple convolutional code so that only certain sequences of subsets are permitted. The sequence is best described by a state transition diagram or "Trellis." Allowed sequences are kept apart, so that maximum distance separation is achieved between branches as seen in Fig. 4.26.

There are many ways to map the coded bits into symbols. Mapping by set partitioning is a technique introduced by Ungerboeck. The technique introduces subsets in the constellation and the Euclidean distance between sequences of signal points in different subsets is significantly increased as seen in Fig. 4.27.

At the receiver, a Viterbi algorithm is used for the combined demodulation and decoding. For each received symbol the distance to the nearest member of each subset is measured. The square of this value serves as the metric in extending the survivor states of the Viterbi algorithm. For each survivor state, not only must the coded bits and accumulated squared distance be stored, but also the uncoded bits corresponding to which member of the subset was the nearest point for each symbol.

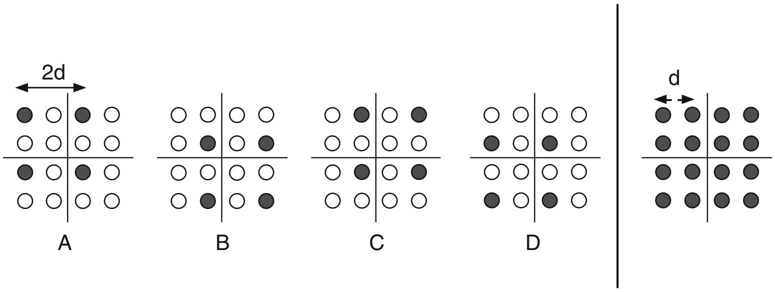


Fig. 4.27 Subsets in the constellation for 16-point

In OFDM, typical Trellis coding is performed over the subcarriers of a symbol. As a result, at the beginning of each symbol the code is started in a known state, and memory is not extended over a greater interval of time.

4.3.6 Turbo Coding

A typical turbo coding[11] includes parallel concatenated convolutional codes where the information bits are coded by two or more recursive systematic convolutional (RSC) codes, which are typically interleaved and optionally punctured as in Fig. 4.28. Let us first look at the encoding process. The components in encoding are:

- The RSC encoder: Why we use recursive format? Convolutional codes operate in feed-forward form such as $(G1,G2) = (1+D^2, 1+D+D^2)$. This structure produces codewords with very low weight, since for instance, a single 1 (...0001000...) gives a codeword equal to generator sequence, and it will propagate through any interleaver as a single 1. This produces larger number of codewords with very low weight. Recursive structure uses division as in $(1, G2/G2) = (1, (1+D+D^2)/(1+D^2))$, which does not change the encoding sequences but changes the mapping of input sequences to output sequences. As a result, a weight-one input gives a codeword of semi-infinite weight, since it diverges from the all-zero path, but never remerges and there will always exists a Trellis path that diverges and remerges later corresponding to a weight-two data sequence.
- The interleaver takes each incoming block of N data bits and rearranges them in a pseudo-random fashion in order to give patterns that has high weight.
- The puncturer periodically deletes the selected bits to reduce coding overhead. Deletion of parity bits is recommended.

[11] "Turbo coding was introduced by Berrou, Glavieux, and Thitimajshima (from ENST Bretagne, France) in their title *Near Shannon Limit error-correcting coding and decoding: Turbo-codes* published in the Proceedings of IEEE International Communications Conference in 1993."

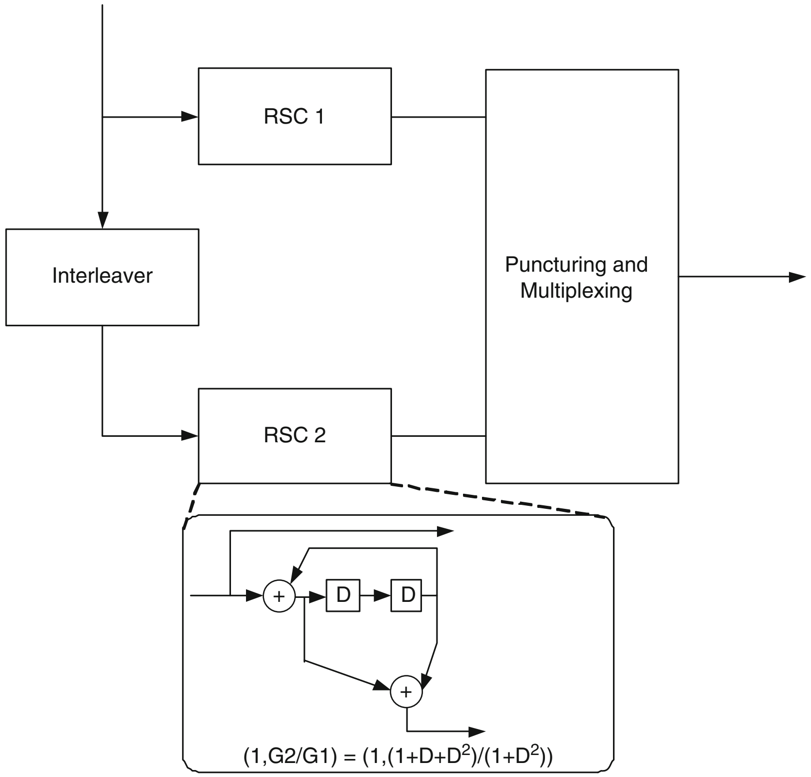


Fig. 4.28 A typical turbo encoder: two identical 1/2 RSC encoder separated by an N-bit interleaver and optional puncturing

The structure and complexity of turbo encoder design is restricted by decoding delay and complexity:

- *Decoding delay* is important for system performance, since significant delay in the system may degrade the system. Increase in the number of parallel coding structure increases the delay.
- *Coding gain* can be increased by designing the system with low SNR but the same BER. However, many other receiver functions such as synchronization and adaptive algorithms require a minimum SNR.

The standard decoding process is iterative as can be seen in Fig. 4.29. We explain two decoding schemes employed by each constituents decoder: Maximum a posteriori probability (MAP) (aka a posteriori probability (APP) or BCJR algorithm) technique and soft input soft output Viterbi algorithm (SOVA). Three different types of soft inputs are available for each decoder:

- The un-coded information symbols,
- The redundant information resulting from first RSC code,
- A priori (extrinsic) information, which is the estimate of the information sequence obtained from the first decoding.

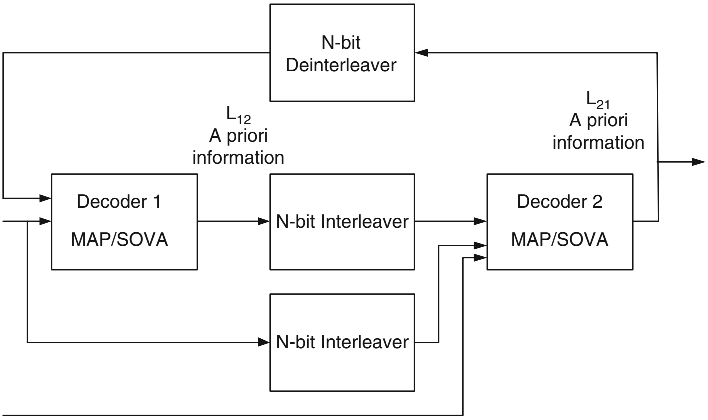


Fig. 4.29 A turbo decoder structure that uses two decoders operating cooperatively

In general, a symbol-by-symbol MAP algorithm is optimal for state estimation of a Markov process. MAP algorithms for turbo decoding calculate the logarithm of the ratio of APP of each information bit being one to the APP of the bit being zero, and the decoder decides $x_k = 1$ if $P(x_k = 1|Y) > P(x_k = 0|Y)$, and it decides $x_k = 0$, otherwise where Y is the received codeword. The log of APP ratio is defined as

$$L(x_k) = \log\left(\frac{P(x_k = 1|Y)}{P(x_k = 0|Y)}\right), \tag{4.13}$$

which translates into

$$L(x_k) = \log\left(\frac{P(Y|x_k = 1)}{P(Y|x_k = 0)}\right) + \log\left(\frac{P(x_k = 1)}{P(x_k = 0)}\right) \tag{4.14}$$

with the second term representing the a priori information. Since $P(x_k{=}1){=}P(x_k{=}0)$ typically, the a priori information is zero for conventional decoders but for iterative decoders, Decoder 1 receives extrinsic information for each x_k from Decoder 2, which serves as a priori information. Similarly Decoder 2 receives extrinsic information from Decoder 1.

The MAP technique is complicated and requires nonlinear operations that make it less attractive for practical purposes. A simplification of MAP algorithm, namely, SOVA leads to a practical suboptimum technique. While performance is slightly inferior to an optimal MAP algorithm, the complexity is significantly less. MAP takes into account all paths by splitting them into two sets, namely, the path that has an information bit one at a particular step and paths that have bit zero at that step and returns the log likelihood ratio of the two.

SOVA considers only the survivor path of the Viterbi algorithm. Therefore, only the survived competing path, which joins the path chosen in the Viterbi algorithm is taken into account for reliability estimation.

While the encoders have a parallel structure, the decoders operate in serial, which results in an asymmetric structure. For example, in the first iteration the first decoder has no a priori information and the same iterative turbo decoding principle can be applied to serial and hybrid concatenated codes as well.

Turbo codes increase data rate without increasing the power of a transmission, or they can be used to decrease the amount of power used to transmit at a certain data rate. Turbo coding shows high error correction performance because of its structure based on interleaving in conjunction with concatenated coding and iterative decoding using (almost) uncorrelated extrinsic information.Turbo coding such as block turbo coding and convolutional turbo coding are included in IEEE 802.16 as supported coding schemes.

4.3.7 LDPC Coding

Turbo codes provide a performance that is very close to the maximum rate dictated by Shannon theorem over a noisy channel as compared with all coding schemes to date. Low-density parity check (LDPC)[12] is an emerging new technique that gets even more closer to Shannon rate with long codewords.[13] LDPC codes are linear block codes that show good block error correcting capability and linear decoding complexity in time.

An LDPC code operates on an **H** matrix containing a low count of ones – hence the name low-density parity-check codes. This is used in encoding in order to derive equations from the **H** matrix to generate parity check bits. Iterative decoding utilizes "soft inputs" along with these equations in order to generate estimates of sent values.

A (n, k) LDPC encoder would have an **H** matrix, which is $m \times n$ in size where $m = n - k$. For instance, a (8,4) LDPC encoder with a code rate of 4/8 might have the following **H** matrix as an example

$$\begin{pmatrix} 01011001 \\ 11100100 \\ 00100111 \\ 10011010 \end{pmatrix}, \tag{4.15}$$

where columns (1–4) are to represent the message and columns (4–8) are to represent the parity bits. It is low density because number of 1s in each row w_r is $\ll m$ and number of 1s in each column w_c is $\ll n$. Also LDPC is regular if w_c is constant for every column and $w_r = w_c(n/m)$ is also constant for every row. Otherwise it

[12] First proposed in 1960 by Robert Gallager in his PhD dissertation and published as "Low-density parity-check codes" in IRE Trans. Information Theory in 1962.

[13] "In 1999, Richardson introduced an irregular LDPC code with code length 1 million. The code is shown to perform within 0.3 dB of the Shannon limit. In 2001, Chung introduced a closer design which is 0.0045 dB away from capacity."

is irregular. There are several mechanisms introduced to construct LDPC codes by Gallager, MacKay,[14] etc. In fact, randomly chosen codes are also possible.

LDPC encoding is similar to systematic block code in, which codeword $(c_0,\ldots,c_n)$ would consist of the message bits $(m_0,\ldots,m_k)$ and some parity check bits as we mentioned earlier. The solution is solving the parity check equations to calculate the missing values

$$\mathbf{H}c^{\mathrm{T}} = 0, \tag{4.16}$$

where this manipulation can be performed with a generator matrix $\mathbf{G}$. $\mathbf{G}$ is found from $\mathbf{H}$, which can be written as follows with Gaussian elimination

$$\mathbf{H} = [\mathbf{P}^{\mathrm{T}} : \mathbf{I}] \tag{4.17}$$

and $\mathbf{G}$ is

$$\mathbf{G} = [\mathbf{I} : \mathbf{P}]. \tag{4.18}$$

Hence, c codeword is found for message word x as follows $c = x\mathbf{G} = [x : x\mathbf{P}]$.

The graphical representation[15] for the same LPDC is given in Fig. 4.30. Graphical representation utilizes variable nodes (v-nodes) and check nodes (c-nodes). The graph has m c-nodes and n v-nodes, where m stands for the number of parity bits. Check node f_i is connected to c_i if h_{ij} of $\mathbf{H}$ is a 1. This is important to understand the decoding. Decoding tries to solve the (n-k) parity check equations of the $\mathbf{H}$ matrix.

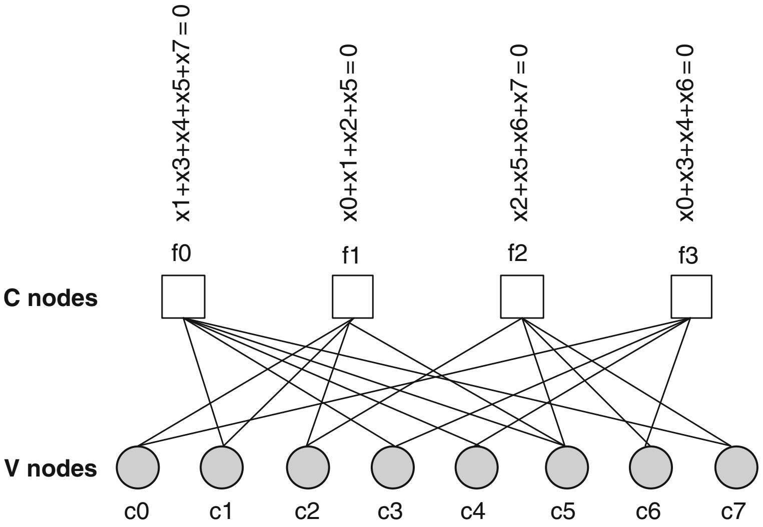


Fig. 4.30 Graphical representation of (8,4) LMDS

[14] "David MacKay, Michael Luby, and others resurrected the LDPC in the mid-1990s."

[15] "Robert Tanner, in 1981 generalized LDPC codes and developed a graphical method of representing these codes, now called Tanner graphs or bipartite graphs..."

There are several algorithms defined to date and the most common ones are *message passing algorithm*, *belief propagation algorithm*, and *sum–product algorithm*.

LDPC decoding is an iterative process where in each round,

Step 1 v-nodes c_i send a message to their c-nodes f_j. In the first step, c_i only has the received bit y_i.

Step 2 c-nodes f_j determine a response to its connected v-nodes. The f_j considers all the messages correct except message$_i$ and calculates a response to c_i. Here, LDPC decoder might find out that the received bits are correct and terminates the decoding if all equations are fulfilled.

Step 3 v-nodes receives these responses from c-nodes and uses this information along with the received bit in order to find out that the originally received bit is correct. Then, sends this information back to c-nodes.

Step 4 go to Step 2.

A simple decoder might use hard decision decoding and would do majority vote in Step 3. An example is depicted in Fig. 4.31 for a codeword $c = [10010101]$. From the figure one can see that the second iteration is enough to detect the correct codeword if bit c_1 is flipped to 1. As you can see in the second step, c-nodes come up with a response for each v-node. If there are even numbers of 1s in all v-nodes except c_i then f_j response to c_i is 0, otherwise 1. Then, in step 3, v-nodes apply majority vote to determine their decision. For example if the originally received bit is 1 for c_0 and messages from check nodes are $f_1 \rightarrow 0$ and $f_3 \rightarrow 1$ then the decision is 1 as well.

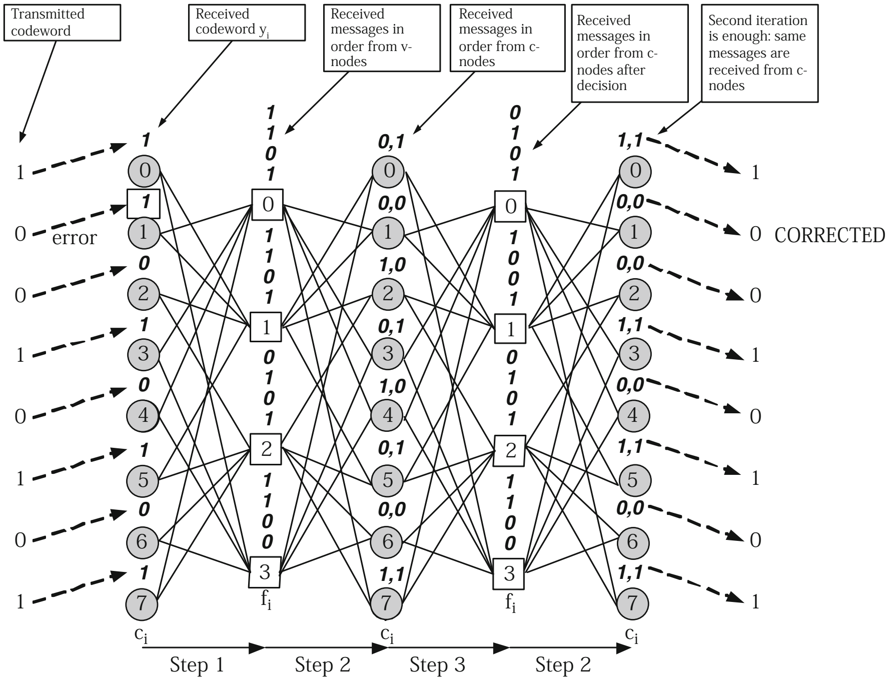


Fig. 4.31 Hard decision decoding for LDPC

Soft decision decoding based on belief propagation introduced by Gallager is preferred since it yields better performance. Belief propagation introduces probabilities as messages that is being passed and these probabilities are used to update the confidence for the bits in the equations.

Let us denote q_{ij} as the message to be passed from v-node c_i to c-node f_j and r_{ji} is the message to be passed from the same c-node to the same v-node. q_{ij} and r_{ij} basically represent the amount of belief in y_i whether it is a "0" or a "1."

In Step 1, all v-nodes send q_{ij} messages. At the first step, $q_{ij}(1) = P_i$ and $q_{ij}(0) = 1 - P_i$, where $P_i = \Pr(c_i = 1|y_i)$. In Step 2, c-nodes calculates the r_{ji}:

$$r_{ji}(0) = \frac{1}{2} + \frac{1}{2}\Pi_{i' \in V_j/i}(1 - 2qi'j(1)), \tag{4.19}$$

where $r_{ji}(1) = 1 - r_{ji}(0)$ and V_j/i all v-nodes except c_i. This is basically the probability that there is an even number of 1s among V_j/i. In Step 3, v-nodes update their responses according to following

$$q_{ij}(0) = K_{ij}(1 - P_i)\Pi_{j' \in C_i/j} r_{j'i}(0) \tag{4.20}$$

$$q_{ij}(1) = K_{ij} P_i \Pi_{j' \in C_i/j} r_{j'i}(1), \tag{4.21}$$

where C_i/j now stands for all c-nodes except f_j and K_{ij} are constants in order to ensure that $q_{ij}(1) + q_{ij}(0) = 1$.

Also at this step, v-nodes update the decision $\hat{c}_i$ with information from every c-node. if the estimated $\hat{c}$ satisfies $\mathbf{H}\hat{c}^{\mathrm{T}} = 0$, then the algorithm terminates. The probabilities for 0 and 1 are found out to be with the following equations

$$Q_i(0) = K_i(1 - P_i)\Pi_{j' \in C_i} r_{j'i}(0) \tag{4.22}$$

$$Q_i(1) = K_i P_i \Pi_{j' \in C_i} r_{j'i}(1), \tag{4.23}$$

where K_i is the constant that makes $Q_i(1) + Q_i(0) = 1$. Hence, $\hat{c}_i$ is 1 if $Q_i(1) > Q_i(0)$, otherwise it is 0. These equations can be modified for log domain as well to change the multiplications into additions.

Unlike turbo coding, LDPC codes can determine when a correct codeword is detected and LDPC decoding based on belief propagation can be simpler than turbo decoding. It can show gains of more than 0.5 dB from a low code rate turbo coding and up to 2 dB from other coding solutions.

4.4 Synchronization

Synchronization is the essential part of the receiver since oscillator impairments and clock differences along with phase noise during demodulation degrade the performance.

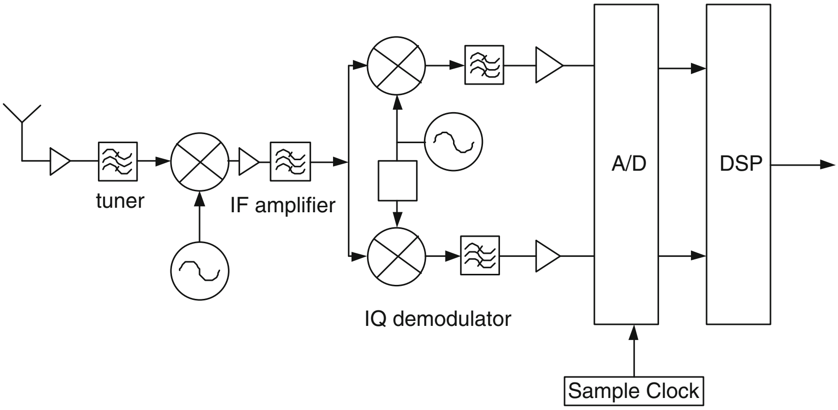


Fig. 4.32 Front end of an OFDM receiver

Let us look at a typical front end of an OFDM receiver, which is depicted in Fig. 4.32. The received signal is first down converted to an IF frequency then to baseband with IQ demodulator. Later, the waveform is converted to digital format by sampling. Receiver first synch with symbol boundary in time domain then it locks to subcarrier frequencies.

Hence, OFDM needs to employ time and frequency synchronization. Time synchronization is to decide for the symbol boundaries. Commonly, a sequence of known symbols-preamble are used to detect the symbol boundaries. It has less sensitivity to timing offset as compared with single-carrier systems, since timing offset does not violate the orthogonality of subcarriers in OFDM system, but causes ISI in single-carrier systems.

Unlike time synchronization, frequency synchronization, which is to estimate the frequency offset in the oscillators in order to align the oscillators in the transmitter and receiver, is essential otherwise ICI occurs, since subcarriers could be shifted from its original position and the receiver may experience nonorthogonal signals. Since the carriers are spaced closely in frequency domain, a small fraction of frequency offset is barely tolerable. Also, practically oscillators do not produce a carrier at exactly one frequency but rather a carrier with random phase noise. This phase noise in time domain corresponds to frequency deviation in frequency domain, thereby causing ICI.

4.4.1 Timing Offset

OFDM is insensitive to timing offset as long as offset is within the guard time. Consequently, no ISI and ICI is guaranteed. On the other hand, optimum symbol detection is important, since any lag in detection may increase the sensitivity to delay spread. Also, timing offset changes the phases of subcarriers but does not violate orthogonality. These phase shifts are estimated during channel estimation if receiver employs coherent receiver. Assuming that there is no ISI, y_k, the received signal, is

$$y_k = \sum_{n=0}^{N-1} x_n e^{j\theta} e^{j2\pi \frac{n}{N} f_s t}\Big|_{t=\frac{t+d}{f_s}}, \tag{4.24}$$

where θ is envelope delay distortion and d is sampling time offset. And $\hat{x}_m$ after FFT is

$$\begin{aligned} \hat{x}_m &= \tfrac{1}{N} \textstyle\sum_{k=0}^{N-1} y_k e^{-j2\pi \frac{m}{N} k} \\ &= \tfrac{1}{N} \textstyle\sum_{n=0}^{N-1} x_n e^{j(\theta + 2\pi \frac{n}{N} d)} \textstyle\sum_{k=0}^{N-1} e^{-j2\pi \frac{k}{N}(n-m)}. \\ &= x_m e^{j(\theta + 2\pi \frac{n}{N} d)}, \quad n = m \\ &= 0, \quad n \neq m. \end{aligned} \tag{4.25}$$

from where we can see that introduced phase offset effects all subcarriers linearly. These phase rotations can be corrected in the channel estimation stage.

4.4.2 Frequency Offset

OFDM is sensitive to frequency offset since it causes ICI, which basically introduces interference from all other subcarriers. If y_k is the received signal with no timing offset and δ_f is the frequency offset, then

$$y_k = \sum_{n=0}^{N-1} x_n e^{j2\pi t(\frac{n}{N} f_s + \delta_f)}\Big|_{t=\frac{k}{f_s}} \tag{4.26}$$

and after FFT we get

$$\hat{x}_m = \Theta(\delta_F) x_m + \sum_{n=0}^{N-1} \sum_{k=0}^{N-1} x_n e^{j2\pi \frac{k}{N}(n-m+\frac{n\delta_f}{f_s})}, \tag{4.27}$$

where $\hat{x}_m$ has an attenuation component ($\Theta(\delta_F)$) as well as interference component from other subcarriers.

4.4.3 Phase Noise

Phase noise occurs during intermediate stages of demodulation, such as between RF and IF. Phase noise is a zero mean random process of deviation of the oscillator's phase. Power spectral density (PSD) of phase noise is normalized with the power of the sine wave, oscillator output. Typical phase noise PSD is shown in Fig. 4.33 and this PSD is high-pass filtered with phase locked loop (PLL) and oscillator is locked for intermediate modulation.

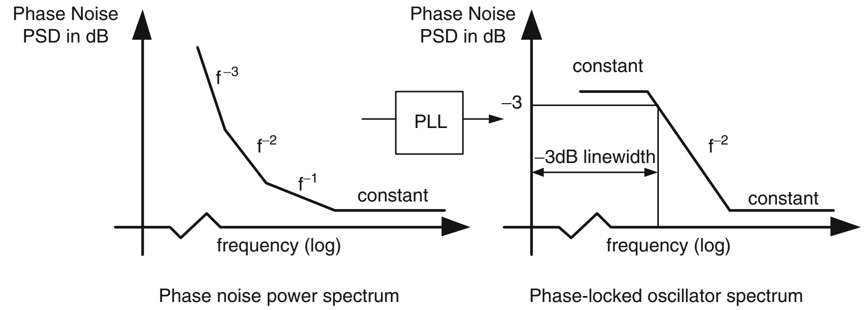


Fig. 4.33 Phase noise power density spectrum

We can follow the same analysis that was used to examine frequency offset, and assume there are no other impairments. Then, the output of the receiver demodulator is

$$\hat{x}_m = \frac{1}{N}\sum_{n=0}^{N-1} x_n \sum_{k=0}^{N-1} e^{j\phi(k)} e^{-j2\pi\frac{k}{N}(n-m)}, \quad m = 0,\ldots,N-1, \qquad (4.28)$$

where $\phi(k)$ is the phase noise for k^{th} subcarrier and can be approximated to $\approx 1 + j\varphi(k)$, thereby

$$\hat{x}_m = x_m - \frac{jx_m}{N}\sum_{k=0}^{N-1}\varphi(k) - \frac{j}{N}\sum_{n=0,n\neq m}^{N-1} x_n \sum_{k=0}^{N-1}\varphi(k) e^{-j2\pi\frac{k}{N}(n-m)}, \qquad (4.29)$$

where we can see that the first term is desired output and the last two terms are two effects of the phase noise: Second term in the equation is a random phase disturbance, which occurs in each symbol. This is known as common phase error (CPE) and can readily be eliminated by measuring the phase variation of a pilot subcarrier and subtracting that rotation from all subcarriers. Third term in the equation is ICI and it is peculiar. The interference may be treated as a Gaussian noise if the different subcarriers are independent. Also decrease in the subcarrier spacing increases the interference.

Figure 4.34 shows the BER performance of phase noise with the following frequency shaping: −65 or −75 dB till 10 KHz offset, slope −20 dB/dec, and −135 dBc noise floor considering the phase-locked spectral shape in Fig. 4.33.

4.4.4 Pilot-Assisted Time/Frequency Synchronization

Time and frequency offset estimator mostly assumes that transmitted data are known at the receiver by transmitting pilot symbols. As a result, symbol timing and carrier frequency offset can be estimated at the receiver. There are also blind methods that uses cylic prefix to estimate synchronization parameters with statistical redundancy.

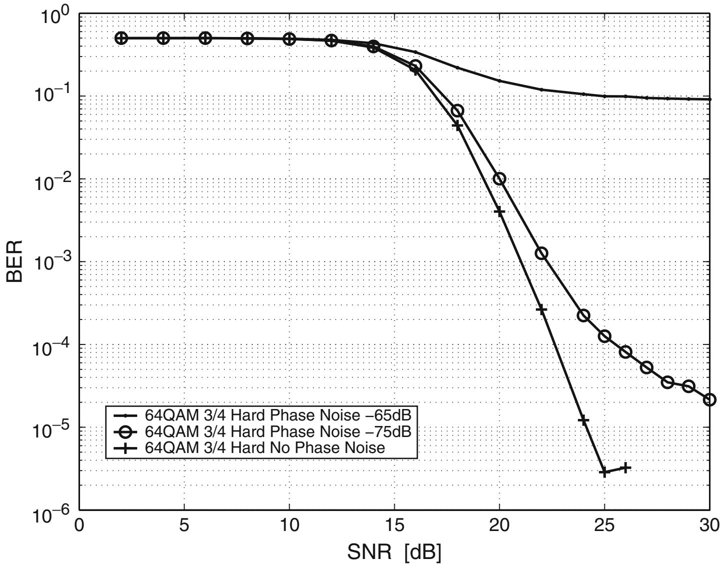


Fig. 4.34 Phase noise (65 dB of 10 KHz)

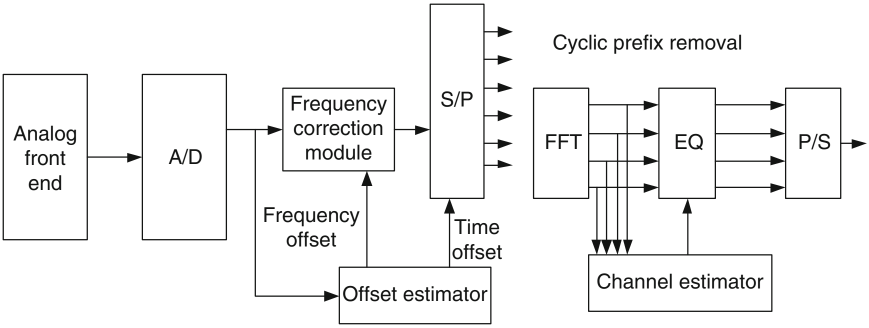


Fig. 4.35 Typical receiver: Frequency correction can be performed by voltage controlled oscillator in the analog front end or digitally by multiplying the received signal in front of FFT with an estimate signal. Timing correction is performed concomitantly with removal of the cylic prefix

Pilot-based methods suit better for high data rate communication, since synchronization should be as quick as possible since blind methods require averaging over a large number of OFDM symbols.

Initially a frame synchronization is performed to detect the start of the frame since typical channel may introduce unknown frequency offset and an unknown time delay. This is performed by correlating the incoming signal with a known preamble. Typical receiver that compensates for time and frequency offsets is shown in Fig. 4.35.

Frame synchronization is achieved by correlating the incoming signal with a known preamble. A matched filter can be used to correlate the input signal with the training signal. From the correlation peaks in the output, symbol timing and frequency offset can be estimated.

4.4.5 Blind Time-Frequency Synchronization

Blind synchronization is pilotless and based on maximum likelihood estimation. The parameters that need to be estimated requires longer observation. As we know, OFDM symbol is cyclically extended, which is basically replicating the tail of the symbol in the guard interval. This cyclostationary property can be leveraged to estimate symbol timing t_o, frequency offset f_o, and symbol width T. The optimum estimator maximizes the following to find t_o, f_o, and T:

$$\Theta(t_o, f_o, T) = \mathrm{Re}\begin{bmatrix} e^{j2\pi f_o T}\sum_{m=1}^{M}\int_{-T_G}^{0} y(t+t_o-mT_s) \\ \times y^*(t+T+t_o-mT_s)\mathrm{d}t \end{bmatrix}, \tag{4.30}$$

where y is received signal, T_G is guard interval length.

4.5 Detection and Channel Estimation

To estimate the transmitted bits at the receiver, channel knowledge is required in addition to the estimates of random phase shift and amplitude change, caused by carrier frequency offset and timing offset. The receiver applies either coherent detection or noncoherent detection to estimate the unknown phase and amplitude changes introduced by multipath fading channel.

The coherent detection of subscribers requires channel estimation. These are done by channel equalizer, which multiplies each subscriber in an OFDM symbol with a complex number.

Noncoherent detection on the other hand does not use any reference values but uses differential modulation where the information is transmitted in difference of the two successive symbols. The receiver uses two adjacent symbols in time or two adjacent subcarriers in frequency to compare one with another to acquire the transmitted symbol.

4.5.1 Coherent Detection

Channel estimation for coherent detection is performed with pilot symbols considering that each subscriber experiences flat fading (Fig. 4.36). This involves sparsely insertion of pilot symbols in a stream of data symbols and measuring the attenuation in these pilot symbols in order to find the channel impulse response for given

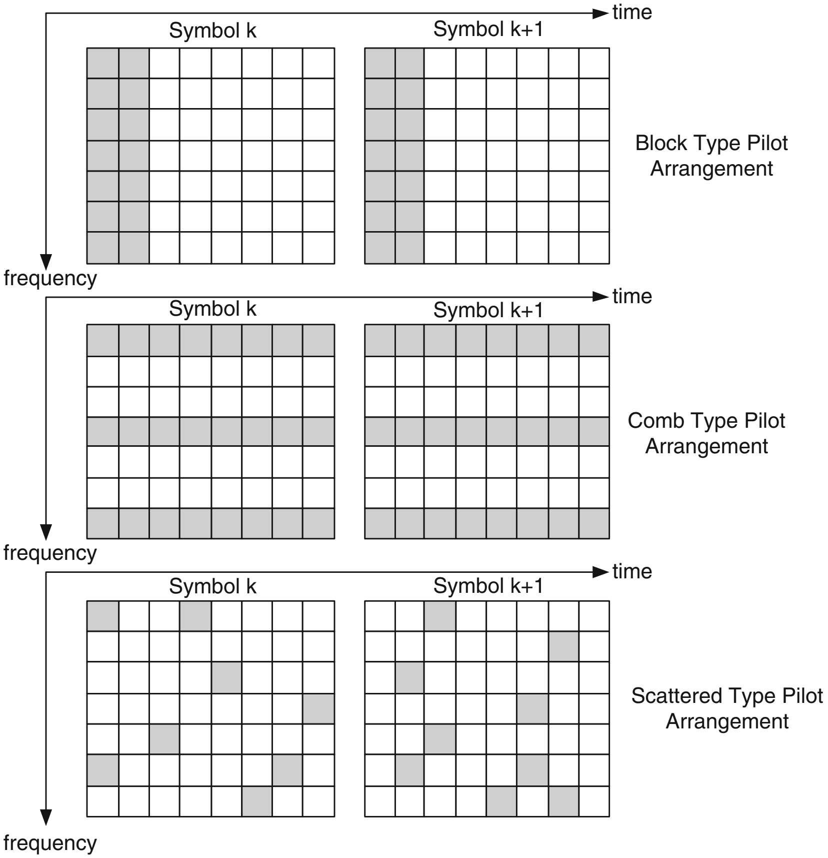


Fig. 4.36 Pilot arrangement for channel estimation

frequency of the pilot and estimating the channel impulse response by interpolating these results to find out the channel components in the data inserted subcarriers.

The coherent detection can be performed by either inserting pilot tones into all of the subcarriers of OFDM symbols with a period in time or inserting pilot tones into each OFDM symbol with a period in frequency. In the case of time-varying channels, the pilot signal should be repeated frequently. The spacing between pilot signals in time and frequency depends on coherence time and bandwidth of the channel. One can reduce the pilot signal overhead by using a pilot signal with a maximum distance of less than the coherence time and coherence bandwidths. Then, by using time and frequency interpolation, the impulse response and frequency response of the channel can be calculated.

Pilot spacing has to follow several requirements in order to strike a balance between channel estimation performance and SNR loss. Smaller pilot spacing in time and frequency would result in a good channel estimation, but the effective SNR for data symbols will be smaller. Channel variation in time and frequency are used to determine the minimum pilot spacing in time and frequency. If B_D is Doppler spread and T_S is delay spread, a suitable choice for pilot spacing in time N_p^t, and in frequency N_p^f, is as follows:

$$N_p^t \approx \frac{1}{B_D T} \qquad N_p^f \approx \frac{1}{\Delta f T_S}\,, \tag{4.31}$$

where Δf is subcarrier bandwidth, T is symbol time.

These requirements lead to several pilot arrangement types in time and frequency. Figure 4.37 illustrates several pilot arrangements. The first one in the figure is a block-type pilot arrangement for slow fading channels. All subcarriers are used once and the estimated channel is used for the coming symbols since the channel is assumed to change very slowly for slow fading channels, and the channel is highly correlated for the consecutive symbols. For example, in wireless LAN systems, block-type pilot arrangement is used, since packet sizes are short enough to assume a constant channel for the duration of the packet.

The second approach is the comb-type pilot arrangement, which interleaves the pilots over time and frequency. Comb-type estimation uses several interpolation techniques to estimate the entire channel. Flat fading channels can be estimated by comb-type pilot arrangement and pilot frequency increases if the channel is frequency selective fading.

Besides interpolation, time and frequency correlation is used for channel estimation. If Doppler effects are kept small by keeping OFDM symbol shorter compared with coherence time of the channel, time correlation between OFDM symbol is high. Moreover, in an ideal OFDM system, if subcarrier spacing is small as compared with the coherence bandwidth of the channel, frequency correlation between the channel components of adjacent subcarriers is high.

Now, assume that the diagonal matrix $\mathbf{X}$ contains the transmitted pilot symbols and vector $\mathbf{y}$ contains the observed output of the FFT:

$$\mathbf{y} = \mathbf{X}\mathbf{h} + \mathbf{n}, \tag{4.32}$$

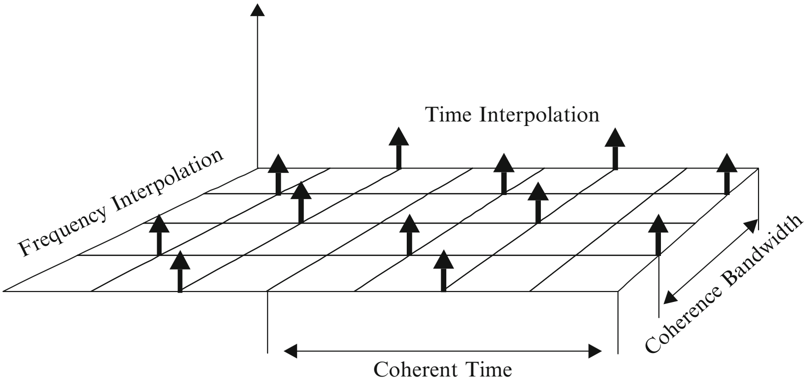


Fig. 4.37 Pilot positioning in time and frequency

where the channel estimation problem is to find the channel estimates $\hat{h}$ as a linear combination of pilot estimates. The least-squares (LS) channel estimation is

$$\mathbf{\hat{h}_{LS}} = \mathbf{X^{-1}y}, \tag{4.33}$$

since LS minimizes $\| \mathbf{y} - \mathbf{X\hat{h}} \|$ for all $\mathbf{\hat{h}}$.

LS operates with received and known transmitted pilot symbols. The frequency correlation can be further exploited with LS in order to minimize $\| \mathbf{\hat{h}} - \mathbf{h} \|^{\mathbf{2}}$ for all possible linear estimators $\mathbf{\hat{h}}$. Then, the optimal linear minimum mean squared error (LMMSE) estimate is

$$\mathbf{\hat{h}_{MMSE}} = \mathbf{A\hat{h}_{LS}}, \tag{4.34}$$

where

$$\begin{aligned} \mathbf{A} &= \mathbf{R_{h\hat{h}_{LS}}} \mathbf{R^{-1}_{\hat{h}_{LS}\hat{h}_{LS}}} \\ &= \mathbf{R_{hh}}(\mathbf{R_{hh}} + \sigma_{\mathbf{n}}^{\mathbf{2}}(\mathbf{XX^{H}})^{-\mathbf{1}})^{-\mathbf{1}}, \end{aligned} \tag{4.35}$$

and $\mathbf{R_{hh}} = \mathbf{E}\{\mathbf{hh^{H}}\}$ is the channel autocorrelation matrix. LMMSE estimator is complex and normally used as a base for designing new estimators.

4.6 Equalization

Equalization is used to combat for intersymbol interference (ISI) and works along with channel estimation. If symbol time is larger than delay spread, then the system suffers from ISI. In OFDM, typically symbol time is extended with guard interval to reduce the ISI. However, equalization is used in the frequency domain to remove amplitude and phase distortions caused by fading channel. Equalizer utilizes channel estimation and continually tracks the channel.

OFDM equalization typically employs a combination of CP with a standard equalization (e.g., linear equalization or decision feedback equalization). Equalization balances ISI mitigation with noise enhancement, since if channel is $H(f)$ and equalizer selects $H_e(f)$ as $1/H(f)$ then the frequency response of noise after equalization $N'(f)$ becomes $.5No/|H(f)|^2$. Notice that for some frequency if there is a spectral null in the channel then noise power is greatly enhanced. In general, linear equalizers cause more noise enhancements than nonlinear equalizers. They can be both implemented using a transversal or lattice structure as seen in Fig. 4.38. The transversal structure is a filter with delay elements and tunable complex weights. The lattice filter has recursive structure, which is more complex than transversal structure but achieves better convergence, numerical stability, and scalability.

Figure 4.39 illustrates time and frequency domain equalization. Frequency domain equalization is used to compensate for channel complex gain at each subcarrier. Notice that ICI is absent in OFDM. Equalization after FFT is equivalent to a convolution of a FIR filter in time domain (residual equalization).

We first talk about time domain equalization methods: linear and nonlinear transversal structure equalizers. Then, we discuss frequency domain equalization.

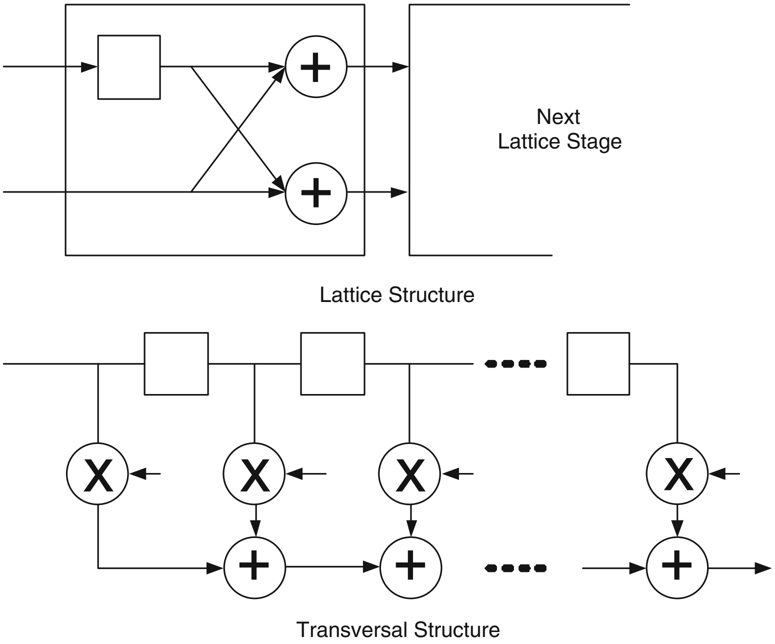


Fig. 4.38 Transversal and lattice equalizer structures

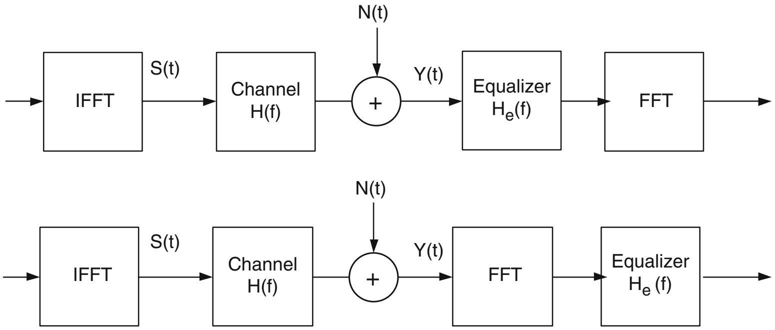


Fig. 4.39 Time and frequency domain equalization

Let us assume a system with a linear equalizers, where pulse shape is compensated with matched filter as seen in Fig. 4.40. If ISI channel is $f(t)$ and transmitted signal is $x(t)$ then received signal $y(t)$ is

$$y(t) = x(t) * f(t) + n(t) = \sum x_k f(t - kT_s) + n(t), \tag{4.36}$$

where $n(t)$ is white noise. When $y(t)$ is sampled with T_s, $y[n] = y(nT_s)$ and $\nu[n] = n(nT_s)$ become

$$y[n] = x_n f[0] + \sum_{k \neq n} x_k f[n-k] + \nu[n], \tag{4.37}$$

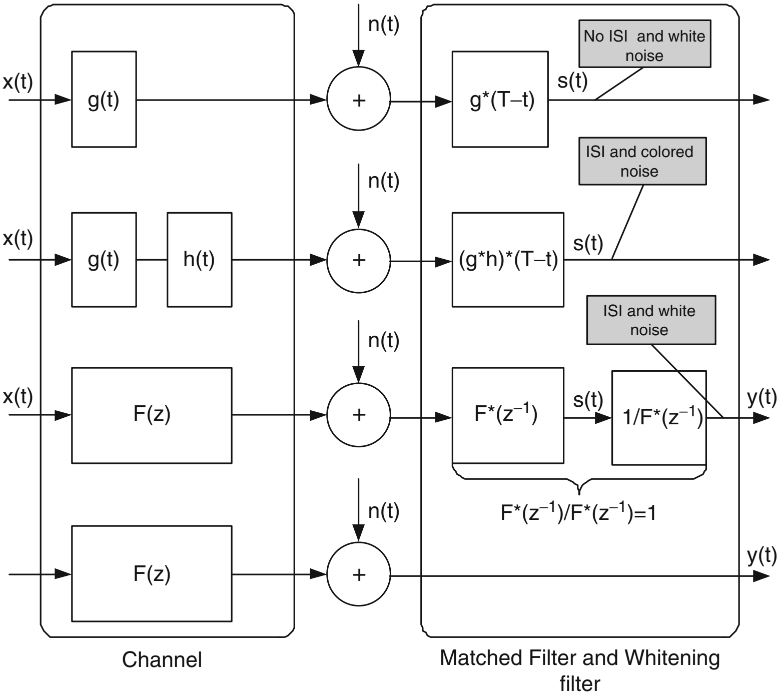


Fig. 4.40 Equalizer model

where the second term stands for ISI and ISI-free[16] communication is achieved if $f[n-k] = 0$ for $k \neq n$.

If there is ISI then the equalizer $F_e(z)$ represented in z-domain is used to reduce the ISI. Linear equalizer is represented as below:

$$F_e(z) = \frac{i=-L}{L} w_i z^{-i}, \tag{4.38}$$

where there are $N = 2L+1$ taps and weights w_i, which are set to reduce the probability of error.

4.6.1 ZF: Zero Forcing Equalizer

ZF equalizer is a linear equalizer and sets $F_e(z)$ to $1/F(z)$, which cancels out the ISI but enhances the noise $N(z)$ by $\frac{1}{|F(z)|^2}$ as seen in Fig. 4.41. This significantly

[16] Folded spectrum is described as

$$F_{\Sigma}(f) = \frac{1}{T_s} \sum_{n=-\infty}^{\infty} F(f + \frac{n}{T_s})$$

and if $F_{\Sigma}(f) = f[0]$, then there is no ISI.

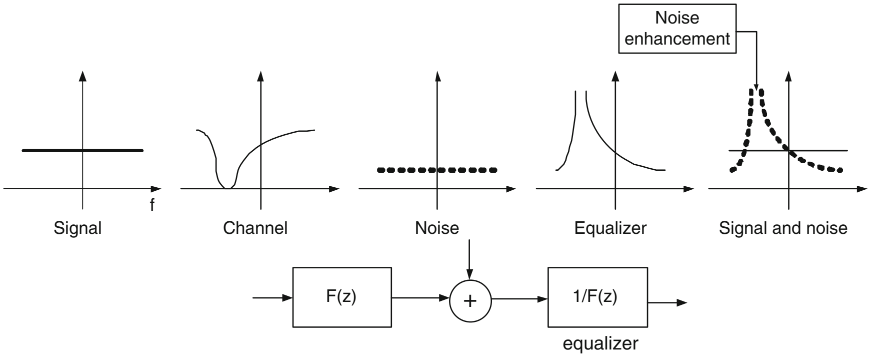


Fig. 4.41 Zero forcing equalizer

increases the noise if there is an attenuation in some frequencies in the channel. Also note that, noise is no longer white but colored, which complicates the detector.

The ZF equalizer determines finite set of coefficients $w_i z_i$ according to $1/F(z)$, of course there are many ways, one technique is setting $w_i = c_i$ in

$$\frac{1}{F(z)} = \sum_{i=-\infty}^{\infty} c_i z^{-i} \tag{4.39}$$

in order to minimize

$$|\frac{1}{F(z)} - \sum_{i=-L}^{L} w_i z^{-i}|^2. \tag{4.40}$$

4.6.2 MMSE: Minimum Mean-Square Error Equalizer

MMSE equalizer is also a linear equalizer and provides a better balance between ISI and noise enhancement. Since it minimizes the expected mean-squared error between transmitted symbol x_k and the symbol detected at the equalizer output $\hat{x}_k$:

$$E[x_k - \hat{x}_k]^2, \tag{4.41}$$

where

$$\hat{x}_k = \sum_{i=-L}^{L} w_i y[k-i] \tag{4.42}$$

and $F_e(z)$ is found to be

$$F_e(z) = \frac{1}{F(z) + N_o}. \tag{4.43}$$

Notice that noise is still colored; when noise is colored, then there is another filter before equalizer, which is called noise whitening filter, which whiten the noise in order to obtain a flat power spectrum.

One can see that if there is no noise then this equals to ZF equalizer, and when there is noise, it clearly shows a balance between reduction in ISI and enhancement in noise as seen in Fig. 4.42.

4.6.3 DFE: Decision Feedback Equalizers

As we see, linear equalizers when taking care of the ISI cause noise coloring. Decision feedback equalizer (DFE), a nonlinear equalizer, detects the symbols and removes the future ISI by subtracting it from the future signal.

DFE equalizer has two stages. Stage 1 is a forward filter to whiten the noise and to produce a response with postcursor ISI only. And stage 2 has the feedback filter to cancel that postcursor ISI as seen in Fig. 4.43.

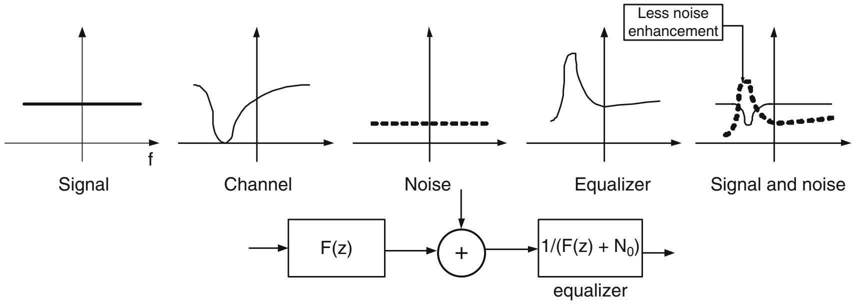


Fig. 4.42 MMSE equalizer

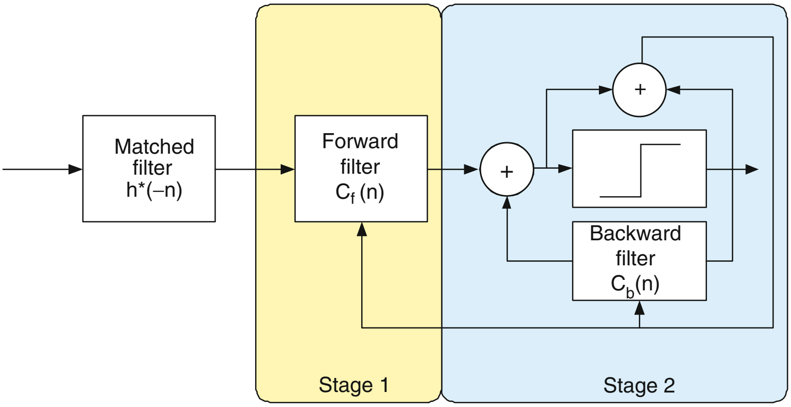


Fig. 4.43 General decision feedback equalizer

The optimum forward filter for a zero-forcing DFE can be considered as a cascade of a matched filter followed by a linear equalizer and a causal whitening filter whose transfer function can be found by spectral factorization of the channel power spectrum.

Let $f(n)$ represent the channel impulse response, hence $f^*(-n)$ is the impulse response of the matched filter. Forward filter $c_f(n)$ performs ISI suppression and noise whitening. The backward filter $c_b(n)$ eliminates the interference of previous symbols, hence $1+c_b(n)$ has to be a monic and causal filter. Therefore, its output should have postcursor intersymbol interference only, namely

$$1+c_b(n) = f(n) * f^*(-n) * c_f(n). \tag{4.44}$$

On the other hand the variance of the noise at the output of the matched filter is

$$\sigma^2 \int_{-\pi/2}^{\pi/2} S_h(e^{j\omega t})\mathrm{d}\omega, \tag{4.45}$$

where S_h is power spectrum of the channel impulse response. To minimize noise power at the input of the slicer, $c_f(n)$ should whiten the noise. Rewriting (4.44) in the frequency domain, the forward filter transfer function is

$$C_f(z) = \frac{1+C_b(z)}{F(z)F^{*^{-1}}(z^{-1})}. \tag{4.46}$$

The forward filter may apply further linear transformation to the input signal to meet some optimality criteria such as minimizing square error or peak distortion. To whiten a stochastic process with spectrum of $\sigma^2 S_h(z)$, it should satisfy the following equation:

$$\sigma^2 S_h(z) \frac{1+C_b(z)}{F(z)F^{*^{-1}}(z^{-1})} \frac{1+C_b^*(z^{*^{-1}})}{F^*(z^{*^{-1}})F(z)} = \sigma^2, \tag{4.47}$$

since

$$S_h(z) = F(z)F^{*^{-1}}(z^{-1}) \tag{4.48}$$

and from (4.47) we have:

$$S_h(z) = (1+C_b(z))(1+C_b^*(z^{*^{-1}})). \tag{4.49}$$

Forcing ISI to be zero before the decision device will cause noise enhancement. Noise enhancement increases the probability of making error, which can easily propagate in such a system. This makes worse than before, since now the system starts to introduce ISI. Also colored noise still requires a complex decision device, which immediately also introduces delay to the system.

In an MSE decision feedback equalizer, the optimum forward filter minimizes the mean-square error before the detector and tolerates some ISI. The same procedure can be followed for the MSE equalizer to obtain

$$S_h^{'}(z) = (1 + C_b(z))(1 + C_b^*(z^{*^{-1}})). \tag{4.50}$$

MSE-DFE reduces the noise enhancement as compared with the ZF-DFE. The performance is higher than other equalizers, but we still have the error propagation problem.

4.6.4 Adaptive Equalizers

Wireless channel is typically time varying $f(t) = f(\tau, t)$. It is better for equalizer to train for the channel periodically to set the equalizer coefficients. Also, equalizer can track and adjust the coefficients with the use of detected data.

Training must be performed within coherence time T_C and depending on the length of training sequence ℓ following equality must hold $(\ell + 1)T_s \geq T_\mathrm{C}$.

Tracking is based on equalizer output bits $\hat{x}_k$ and threshold detector output bits $\hat{\hat{x}}_k$, since $\hat{x}_k$ is round to the nearest constellation point via threshold detector. If there is an error, then that error is used to adjust the coefficients of $H_e(z)$ to minimize $\hat{x}_k$ and $\hat{\hat{x}}_k$ via MMSE procedure.

There are available techniques for training and tracking. Performance metrics are number of multiplication for N tap, complexity, training convergence time, and tracking performance. MMSE (number of multiplication is N^2 to N^3) is the most complex with fast training and tracking and least mean square (LMS) $(2N + 1)$ is least complex but with slow training and tracking. LMS updates the tap weight vector linearly with a step size $(w_i(k+1) = w_i(k) + \Delta \varepsilon_k)$, where Δ dictates the convergence speed and stability. Other techniques such as root least square (RLS) $(1.5N^2 + 4.5N)$ and fast Kalman DFE $(20N + 5)$ lie in between the two.

4.6.5 MLSE: Maximum Likelihood Sequence Estimation

Maximum likelihood sequence estimation (MLSE) is a nonlinear estimation technique that replaces the equalizing filter with a MLSE estimation as seen in Fig. 4.44.

MLSE compares the received noisy sequence $\{x_j\}$ with all possible noise-free received signal $y(t)$ and select the closest one. MLSE avoids noise enhancement. MLSE is optimal but very complex to implement, since for length n, there are 2^n different noise free sequences to compare.

4.6.6 Viterbi Equalizer

The Viterbi-equalizer seen in Fig. 4.44 reduces the complexity and minimizes the probability of detecting the wrong sequence of symbols. Fig. 4.45 depicts an

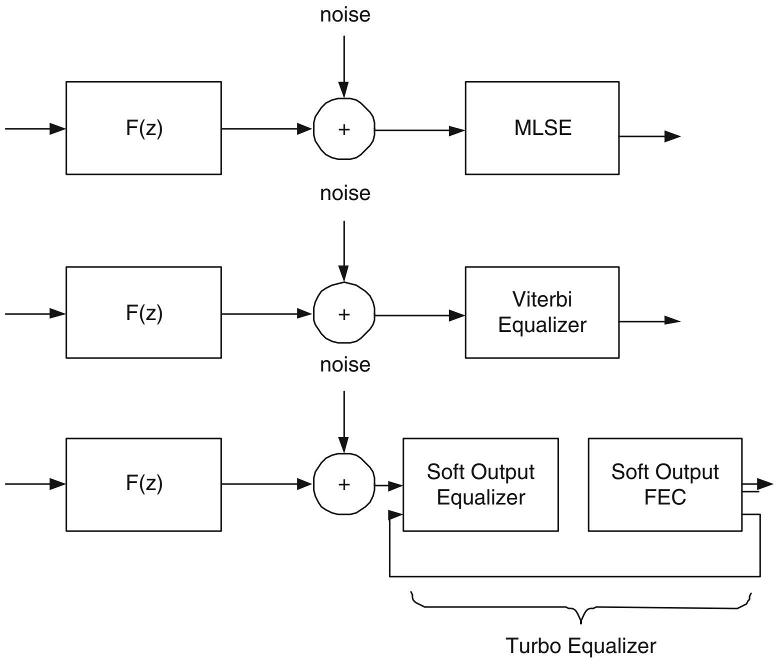
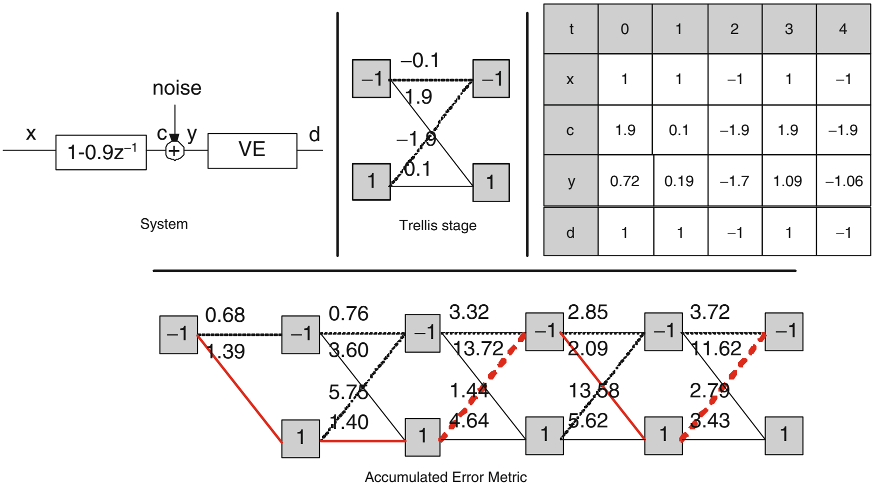


Fig. 4.44 Other equalizers



t	0	1	2	3	4
x	1	1	−1	1	−1
c	1.9	0.1	−1.9	1.9	−1.9
y	0.72	0.19	−1.7	1.09	−1.06
d	1	1	−1	1	−1

Fig. 4.45 An example for Viterbi equalizer

example for Viterbi equalizer. Notice that Viterbi equalizer operates with soft decision, where error metric is obtained by Euclidean distance. For instance at $t = 0$, Euclidean distance 1.39 for output one (from state -1 to state 1) is obtained by

$|1.9-0.72|^2$ and for output zero (from state -1 to state -1), 0.68 is obtained by $|0.72-(-0.1)|^2$. In each branch, Euclidean distance is added to the previous accumulated error metric and at the end, depending on the minimum accumulated error metric, the equalizer traces back to find the correct path and output.

For transmitted sequences of length n over a length $L+1$ channel, it reduces the brute-force maximum-likelihood detection complexity of 2^n comparisons to n stages of 2^L comparisons through elimination of Trellis paths where $L \ll n$. If the length of the channel is high, then its complexity increases as well.

4.6.7 Turbo Equalizer

Turbo equalizer seen in Fig. 4.44 is an iterative equalizer that utilizes a maximum a posteriori (MAP) equalizer and a decoder. The MAP equalizer considers a posteriori probability (APP) of the transmitted symbol with the past channel outputs. The decoder computes log likelihood ratio (LLR) with the transmitted symbol and past channel outputs. After some iteration, the turbo equalizer converges to estimate the transmitted symbol.

4.6.8 Equalization in OFDM

We know that OFDM systems use a cyclic prefix (CP) – guard interval – that is inserted at the beginning of each symbol. This transforms the linear convolution of data and channel into a circular one. As long as the CP is beyond delay spread, ISI is avoided. Otherwise, ISI is present and can degrade performance. As a result, unlike a single carrier system, which utilizes equalizer to minimize ISI, an equalizer can be utilized only to limit the length of ISI, since it is adequate to reduce it to a time span, which is less than the length of CP.

Of course, an equalizer is also utilized to reduce the required CP, since CP reduces the efficiency of the system. Reduction of the transmission efficiency is by a factor $N/(N+N_{\mathrm{CP}})$ and more severe when the transmitted symbol rate is higher, because it requires longer CP. Effective channel impulse response (EIR) can be shorter than the selected CP duration with a time domain equalizer before FFT demodulation at the receiver.

Block diagram of a time domain equalizer for OFDM differentiates from DFE equalizer as seen in Fig. 4.46. The feedback filter is not in the loop and there is no requirement to be monic or causal, since truncating the impulse response does not require a monic feedback filter. The objective is to shorten the sampled CIR of length N_h, $h=[h_0,\ldots,h_{N_h-1}]^{\mathrm{T}}$ to an EIR having significant samples for a length N_b, where $N_f < N_h$, with the use of a time domain equalizer of length N_f, $f=[f_0,\ldots,f_{N_f-1}]^{\mathrm{T}}$. The error term is

$$e_k = f^{\mathrm{T}} * r - b^{\mathrm{T}} * x, \tag{4.51}$$

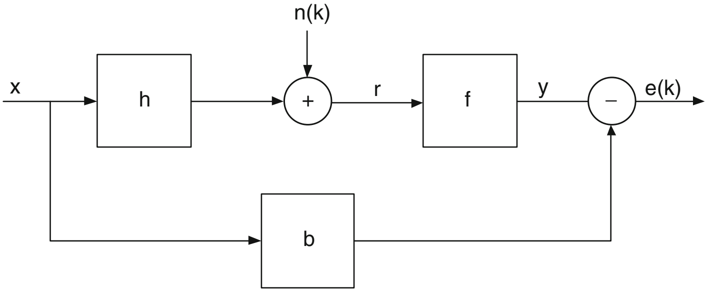


Fig. 4.46 Time domain equalizer configuration

where x and r are vectors of training and received samples respectively. The squared error is given by

$$E\{|e(k)|^2\} = f^{\mathrm{T}} R_{rr} f^* + b^{\mathrm{T}} R_{xx} b^* - f^{\mathrm{T}} R_{rx} b^* - b^T R_{rx} f^*, \tag{4.52}$$

where R_{rr}, R_{xx}, and R_{rx} are the corresponding correlation matrices of r and x. The optimal f can be obtained by MMSE, which is given by

$$\mathrm{d}(E\{|e(k)|^2\})/\mathrm{d}f = 0 \tag{4.53}$$

and this leads to

$$f = R_{rr}^{-1} R_{rx} b. \tag{4.54}$$

We can solve for b by substituting the above relation in (4.52) and

$$E\{|e(k)|^2\} = b^{\mathrm{T}}(R_{xx} - R_{rx}^{\mathrm{T}} R_{rr}^{-1} R_{rx}) b^* = b^{\mathrm{T}} \theta b^*, \tag{4.55}$$

where b is then found as the eigenvector, which corresponds to the smallest eigenvalue of the matrix θ.

In general, we can use adaptive techniques to find near optimum equalizer coefficients. A large number of taps prevents us from using MMSE type of algorithms. On the other hand, a wide spread of eigenvalues of the input signal covariance matrix can slow down convergence of the equalizer. A different iterative technique for the equalizer tap adjustment is based on the steepest gradient methods such as LMS as explained below. Hence,

$$\begin{aligned} f^{k+1} &= f^k - \Delta_1 e(k) r^* \\ b^{k+1} &= b^k - \Delta_2 e(k) x^*, \end{aligned} \tag{4.56}$$

where Δ are the LMS convergence control parameters and equalizer taps are adjusted during the training sequence where transmitted sequence is assumed to be known and the same tap values are preserved during data mode.

The estimation of channel impulse response can help in choosing the optimum window location. As explained previously, b is not necessarily causal; therefore, delay parameter plays an important role in the performance of the equalizer. Other important parameters are the lengths of the two filters f and b.

Typically, the channel impulse response is selected as an ARMA model of the form

$$\frac{A(z)}{1+B(z)} \tag{4.57}$$

and classic decision feedback equalizers; an ARMA equalizer can have a feedback filter with noncausal transfer function, i.e., the general error term is:

$$f * r_k - b * x_{k-d} \tag{4.58}$$

the delay unit d has significant effect on overall performance. The number of taps M of filter b is approximately the same as the prefix length and number of taps for the forward filter is $N \leq M$.

The delay unit should be chosen properly to capture the most significant window of the channel impulse response (CIR). Brute-force trial and error is one option. A more intelligent technique uses the estimate of the impulse response to pick a proper starting point where the energy of the impulse response (unit energy constraint UEC) is above a threshold. The threshold should be the ratio of the tap power to the overall window power. Since the length of the window function and the forward equalizer (FEQ) should be as short as possible, the proper choice of starting point is of significant importance.

Otherwise, AWGN is amplified by the forward equalizer, since FEQ tap coefficients are calculated as the inverse of the EIR. Consequently the subchannels that fall in the nulls are severely degraded because of the low SNR.

Besides robustness against ISI, OFDM is also robust against ICI. But if ICI is not completely avoided, then orthogonality of adjacent subcarriers are not preserved in the frequency domain.

Once ICI free transmission is assured that orthogonality of the subcarriers is maintained, possibly through the use of the cylic prefix and time domain equalization as previously discussed, then the frequency domain equalization of an OFDM signal is an extremely simple process. This is certainly one of the key advantages of OFDM.

After demodulation, the subcarriers will be subjected to different losses and phase shifts, but there will be no interaction among them. Frequency domain equalization therefore consists solely of separate adjustments of subcarrier gain and phase, or equivalently of adjusting the individual decision regions. In the case where the constellations consist of equal amplitude points, as in PSK, this equalization becomes even simpler in that only phase needs to be corrected for each subcarrier, because amplitude has no effect on decisions.

A simplified picture of the place of frequency domain equalization is shown in Fig. 4.47, where the equalizer consists of the set of complex multipliers, $\{A\}$, one for each subcarrier.

Here the linear channel transfer function $F(f)$ includes the channel, the transmit and receive filters, and any time domain equalization if present. $F(f)$ is assumed

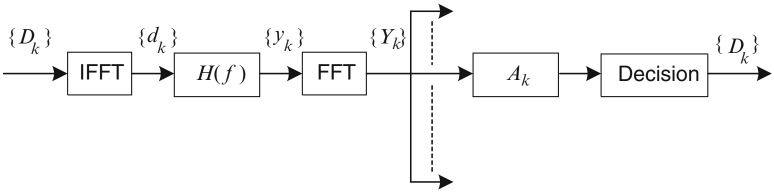


Fig. 4.47 An OFDM system with frequency domain equalization

bandlimited to less than N/T for a complex channel. Cyclic extension is not shown although it is almost certain to be present. The following analysis assumes that both amplitude and phase need to be corrected, and that equalization consists of multiplying each demodulated component by a quantity such that a fixed set of decision regions may be used.

The signal presented to the demodulator is

$$y(t) = \sum_{n=0}^{N-1} d_n h(t - n\frac{T}{N}). \tag{4.59}$$

The kth output of the demodulator is then

$$y_k(f) = x_k F_k, \quad k = 0, 1, \ldots, N-1, \tag{4.60}$$

where the F_k are samples of $F(f)$

$$F_k = h\left(\frac{k}{T}\right). \tag{4.61}$$

Thus each output is equal to its associated input data symbol multiplied by a complex quantity, which differs among the outputs, but are uncoupled. Equalization at its simplest then consists of setting the multipliers to $1/F_k$ for each nonzero channel.

The above approach is optimum in every sense under high signal-to-noise conditions. It also produces minimum probability of error at any noise level, and is an unbiased estimator of the input data x_k. However if the criterion to be optimized is minimum mean-square error particularly, then the optimum multipliers are modified to

$$A_k = \frac{1}{F_k} \frac{1}{1 + \frac{\sigma_k^2}{|x_k F_k|^2}}, \tag{4.62}$$

where σ_k^2 is the noise power in the demodulated subchannel. However, this value produces a biased estimator and does not minimize error probability.

As a practical implementation issue, for variable amplitude constellations, it is frequently desirable to have a fixed grid of decision regions. The A_ks can then be scaled in amplitude such that the separation of constellation points is constant, since a shift τ in timing phase is equivalent to a phase shift

$$F_k = e^{j2\pi \frac{k}{T}\tau} \tag{4.63}$$

and frequency domain equalization readily corrects for such timing shift.

In principle frequency domain equalization could be employed when orthogonality is lost because of interference among OFDM symbols. In this case, rather than a simple multiplier per subchannel, a matrix multiplication would be required. This approach is bound to require more computational load than the combination of time and frequency domain equalizers.

4.6.9 Time and Frequency Domain Equalization

During system initialization, any time domain equalizer must be adjusted before frequency domain equalization is performed. Then any periodic test signal with full frequency content, such as a repeated segment of a PN sequence without cylic prefix, may be used to adapt the frequency domain equalizer.

An interesting interpretation of time domain and frequency domain equalizer can be obtained by studying their role in channel distortion compensation. As discussed earlier, a typical channel model for OFDM and channels with long impulse response is an ARMA model of the form in (4.58) where time domain equalization shortens the impulse response to a tolerable level for an OFDM system. Mathematically, it is equivalent to compensating the AR part of the channel impulse response $1/(1+B(z))$. So after successful time domain equalization, the equivalent impulse response of the channel is reduced to a FIR filter of short duration $A(z)$. Since it does not violate orthogonality of subcarriers, we can remove its effect after the FFT by frequency domain equalization as seen in Fig. 4.48.

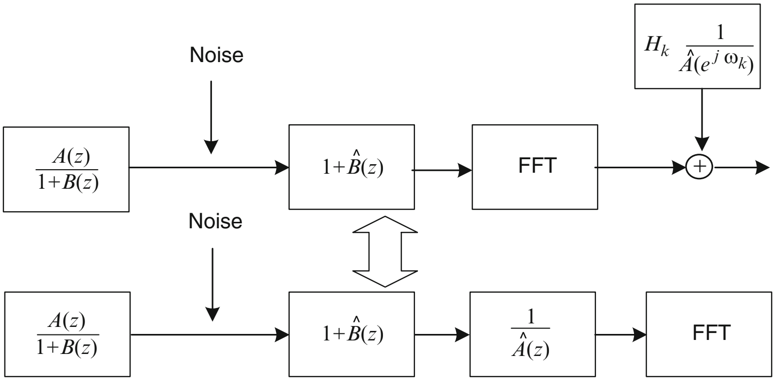


Fig. 4.48 Time and frequency domain equalization

4.7 Peak-to-Average Power Ratio and Clipping

One of the significant drawback of OFDM system is the possibility to experience large peaks since the signal shows a random variable characteristic since it is sum of N independent complex random variables. These different carriers may all line up in phase at some instant and consequently produce a high peak, which is quantified by peak-to-average-power ratio (PAPR).

This distorts the transmitted signal if the transmitter contains nonlinear components such as power amplifiers (PAs). Since PA is forced to operate in the nonlinear region. The nonlinear effects may cause in-band or out-of-band distortion to signals such as spectral spreading, intermodulation, or change the signal constellation. Out-of-band distortion is detrimental even if the in-band distortion is tolerable. To have distortionless transmission, the PAs require a backoff, which is approximately equal to the PAPR. This decreases the efficiency for amplifiers and increases the cost. High PAPR also requires high range and precision for the analog-to-digital converter (ADC) and digital-to-analog converter (DAC), as a result, reducing the PAPR of practical interest.

4.7.1 What is PAPR?

Figure 4.49 depicts a PA. One can see the nonlinear behavior of the PA. It is desired to operate the PA in the linear region. To avoid the high peaks, average input power may be decreased. Operating region of the PA is called input back-off and the resultant signal is guaranteed to be in output back-off range. High input backoff reduces the power efficiency and would mandate the cost of the PA higher, since input backoff is usually greater than or equal to the PAPR of the signal. Ideally, the average and peak values should be as close as can be in order to maximize the efficiency[17] of the PA. PAPR mitigation relaxes the PA backoff requirements as well as the high resolution requirements on ADC and DAC.

PAPR mitigation may fall into three categories: signal distortion, coding, and scrambling. Signal distortion basically distorts the signal around peaks with either clipping or peak windowing or peak cancelation. Coding utilizes forward error correction coding schemes to achieve signals with low PAPR. Scrambling also similar to coding utilizes scrambling sequences to achieve low PAPR. Let us first analyze the PAPR to get a better insight on the mitigation techniques, which we explain later in this section.

[17] "The theoretical relationship between PAPR and transmit power efficiency is given by

$$\eta = \eta_{\max} 10^{\frac{\mathrm{PAPR}}{20}}, \tag{4.64}$$

where η is power efficiency and $\eta_{\max}$ is maximum power efficiency. $\eta_{\max}$ is 50% and 78.5% for Class A and Class B power amplifiers."

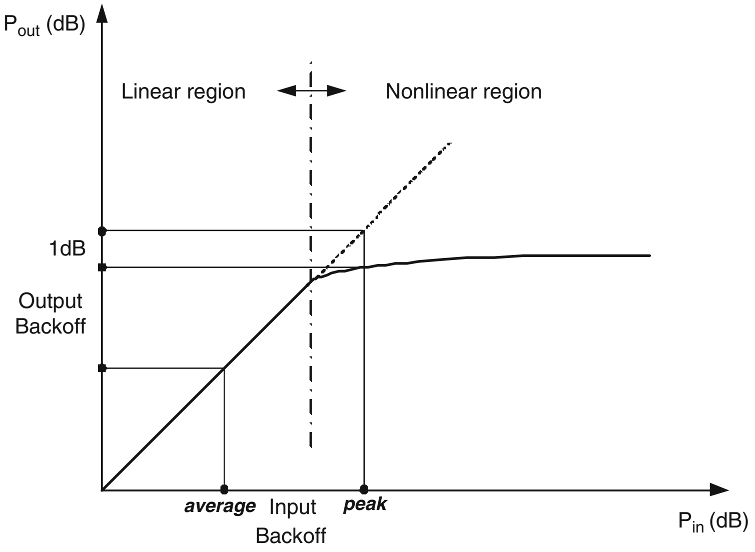


Fig. 4.49 Power amplifier 1 dB compression point: It is desirable to make power amplifier remain linear over an amplitude range that includes the peak amplitudes. Parameters to describe the nonlinearities of the PAs include amplitude modulation/amplitude modulation (AM/AM) distortion, amplitude modulation/phase modulation (AM/PM) distortion, 1 dB compression point (P1dB), and 3rd order interception point (IP3)

If $\mathbf{X}$ is data vector of length N, time domain vector in the transmitter is $\mathbf{x} = [x_0, \ldots, x_{N-1}] = \mathrm{IDFT}(\mathbf{X})$ and the PAPR is then defined to be

$$\mathrm{PAPR}(\mathbf{x}) = \frac{||\mathbf{x}||_\infty^2}{E(||\mathbf{x}||_2^2)/N}, \tag{4.65}$$

where $E(.)$ denotes expectation. $||.||_\infty$ and $||.||_2$ represent the ∞-norm and 2-norm respectively. Therefore, $||.||_2^2 = \sigma^2$ denotes the average (RMS) power. When N is large, the output time vector converges to Gaussian distribution due to central limit theorem. Hence, the probability that the PAPR is above a threshold is written as

$$\Pr\{\mathrm{PAPR} > \lambda\} = (1 - (1 - e^{-\lambda})^N). \tag{4.66}$$

This is plotted for different values of N, and as it can be seen from Fig. 4.50, the system is more susceptible to PAPR when subcarrier size increases. For a baseband OFDM signal with N subcarriers, PAPR may be as large as $N^2/N = N$ for PSK modulation if N subchannels add coherently. Reduction of subcarrier is one way to reduce PAPR but not efficient. Also from the figure, we can infer that high PAPR does not occur often. Considering these infrequent large peaks, a common approach is to perform clipping in order to mitigate the PAPR. These peaks are removed at a cost of self-interference and bandwidth regrowth. As long as these impairments are kept as small as can be, clipping is a powerful and simple technique to employ.

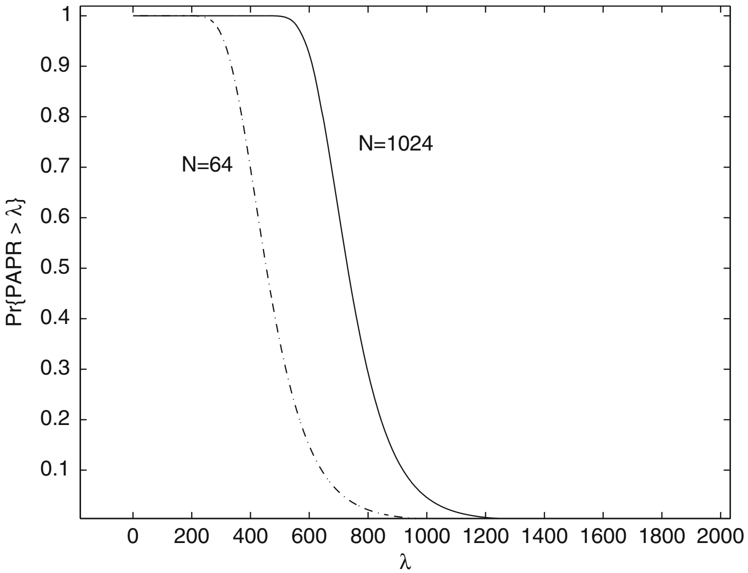


Fig. 4.50 Cumulative distribution function for PAPR

Also it has been proven that the absolute peak presented above is not a good measure to define the "peak" of the signal power. A good measure defines the "peak" as the level that the probability of crossings that level is negligible. As a result, clipping would occur whenever the signal exceeds this "peak," which is typically defined as *m*-fold of average RMS power.

4.7.2 Clipping

Clipping is a nonlinear process and limits the amplitude at some desired maximum level. This simple mechanism introduces the following impairments: self-interference and out-of-band leakage. There are two prong ways to analyze the distortion caused by clipping: additive Gaussian noise or sporadic impulsive noise. They differ with respect to clip level since if clip level is low then clipping events are high, which tends to a Gaussian-like noise. If clip level is high then the clipping events are sporadic. Then the clipping forms a kind of impulsive noise rather than a continual background noise.

Let us define clipping system first as illustrated in Fig. 4.51. If input $x(t)$ is a multicarrier signal, output $y(t)$ after clipper is as follows

$$y = h(x) = \begin{cases} -l\ x \leq l \\ x \quad |x| < l, \\ l \quad x \geq l \end{cases} \tag{4.67}$$

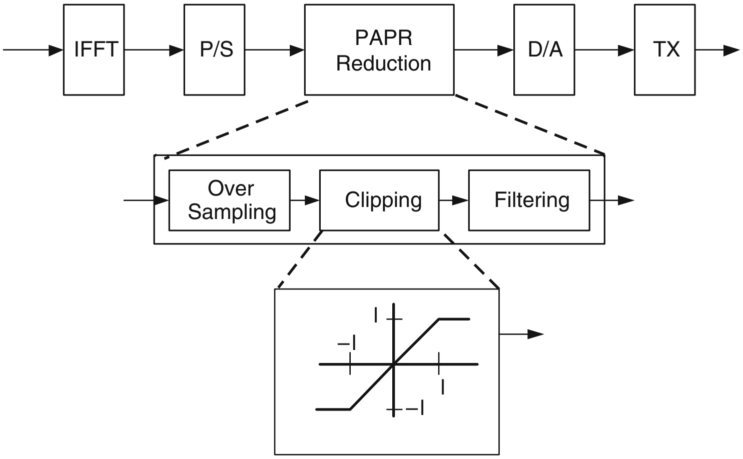


Fig. 4.51 Clipping and Filtering

where $h(.)$ is the nonlinear transfer function of clipping. Bussbang's theorem may decompose the $y(t)$ in two uncorrelated signal components

$$y(t) = \kappa x(t) + c(t), \tag{4.68}$$

where $\kappa \approx 1$ for $l \gg 1$. A clipping scenario is depicted in Fig. 4.52, where we can see that the clip level determines the frequency of occurrences. We first present the bit error rate analysis (BER) because of in-band distortion and then talk about the out-of-band distortion.

4.7.2.1 In-Band Distortion

We start with impulsive noise model and compare this with Gaussian model. The clip level crossings is elaborated as Poisson process in the literature. The rate of the Poisson process is determined by the power spectral density of the signal. Hence, the rate of the Poisson process is

$$\lambda = \frac{f_0}{\sqrt{3}} e^{-\frac{l^2}{2}}, \tag{4.69}$$

where $f_0 = N/T$ (T is OFDM symbol duration) stands for the rectangular region in the power spectrum of the $x(t)$ and $P(C)$, probability of clip, is $2\lambda T$ for double-sided clipping. Also the length of the signal, which the signal stays above the clip level, is found to be (asymptotically) Rayleigh distributed where the expected value for the duration of a clip is given by

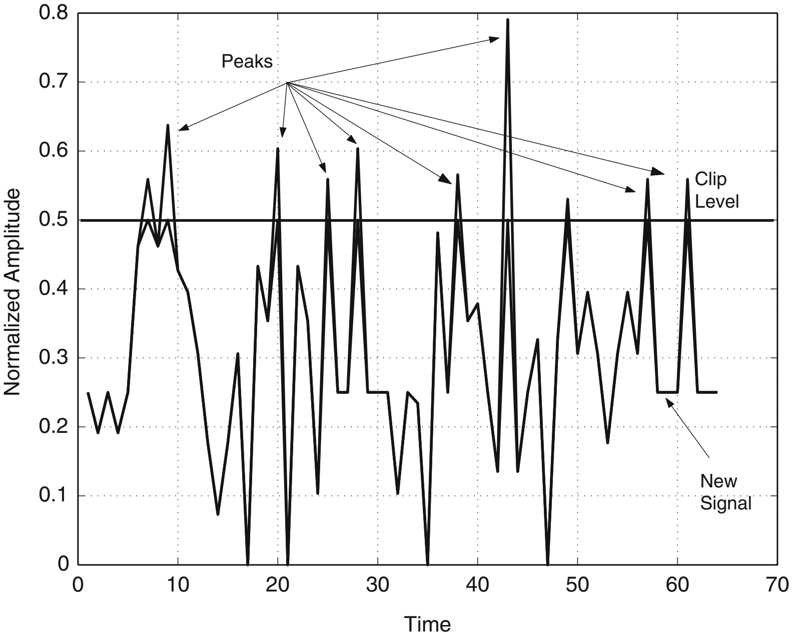


Fig. 4.52 Clipping

$$\frac{1}{\sqrt{\frac{2\pi}{3} f_0 l}} \tag{4.70}$$

from where we can deduct that increase in the clip level (l) decreases the crossing rate and duration. On the other hand, increase in the subcarriers (N) increases the rate and duration. The BER analysis requires to find $\Pr(\text{error}|C)$, the error probability given that there is a clipping $P(C)$. Then this is added to $\Pr(\text{error}|C^c)$, the error probability given that there is no clipping $P(C^c)$. As a result, the error probability $\Pr(\text{error})$ is

$$\Pr(\text{error}) = \Pr(\text{error}|C)\Pr(C) + \Pr(\text{error}|C^c)\Pr(C^c). \tag{4.71}$$

$\Pr(\text{error}|\text{C})$ is found to be $4\frac{4(L-1)}{L} Q\left(\left[\frac{3\pi l^2}{\sqrt{8(L^2-1)}}\right]^{1/3}\right)$ with a square constellation of L^2 points assumption.[18] This is an upper bound and it is interesting to note that error due to clipping varies across subcarriers, the lower subcarriers dominate the overall error more than the others. The overall probability of symbol error is upper bounded by

[18]
$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} e^{-\frac{u^2}{2}} \,\mathrm{d}u. \tag{4.72}$$

$$Pr(error) = \frac{8N(L-1)}{\sqrt{3}L} e^{-\frac{l^2}{2}} Q\left(\left[\frac{3\pi l^2}{\sqrt{8(L^2-1}}\right]^{1/3}\right). \quad (4.73)$$

On the other hand, the BER for Gaussian model, which holds only if the clip level is low enough to permit more frequent occurrence per OFDM symbol is given by

$$\Pr(\text{error}) = \frac{4N(L-1)}{L} Q\left(\frac{\sqrt{3}}{\sigma_c\sqrt{(L^2-1}}\right), \quad (4.74)$$

where σ_c^2 is the power of the clipped portion and if signal power is normalized to unity, it is given by

$$\sigma_c^2 = -\sqrt{\frac{2}{\pi}} l e^{-\frac{l^2}{2}} + 2(1+l^2)Q(l). \quad (4.75)$$

Figure 4.53 plots these probabilities in addition to the BER performance in the receiver when AWGN and Rayleigh fading is present in the channel. One can see from the figure that if clipping level is very low, both models does not accurately model, since almost all signals are clipped below clip level 2.

Also around clip level 2, impulsive model is not that accurate as Gaussian model, since the impulses are considered to be concentrated not spread over time. Hence, this leads to greater probability of error as compared with real implementation. If clip level increases as can be seen that Gaussian noise only considers reduction

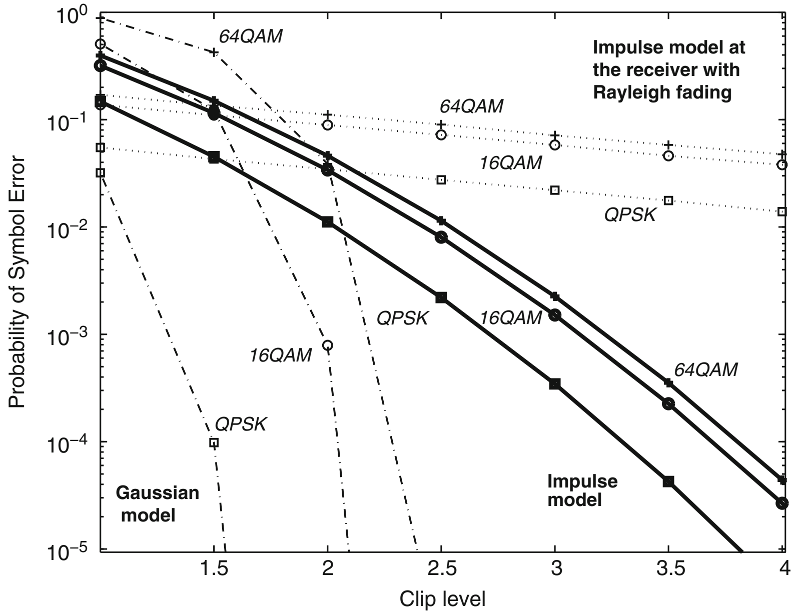


Fig. 4.53 In-band distortion – normalized with N

in the noise power but impulsive noise model considers spikes, which cause errors since small noise underestimates the error probability by several orders of magnitude. If we compare these with the BER performance in the receiver, illustrated in the same figure, we see that the performance is far worse than the one that obtained in the absence of fading, since now the channel introduces variable amplitude z and corresponding overall error probability is given by

$$\Pr(\text{error}) = \frac{4\pi N(L-1)}{L}\int_0^\infty z e^{-\frac{\pi z^2}{4}} Q\left(\left[\frac{3\pi l^2 z^2}{\sqrt{8(L^2-1)}}\right]^{1/3}\right). \tag{4.76}$$

4.7.2.2 Out-of-Band Distortion

Out-of-band distortion (or spectral reqrowth) caused by clipping widens the bandwidth. As can be seen from Fig. 4.54, spectral regrowth with decrease in clipping level directly increases adjacent channel interference and reduces the energy of in-band signal. This is important since spectrum leakage is subject to regulatory limits and directly determines the filtering requirements. In some applications, even if in-band distortion is acceptable or insignificant, out-of-band distortion is intolerable.

The PSD of spectral regrowth is given by

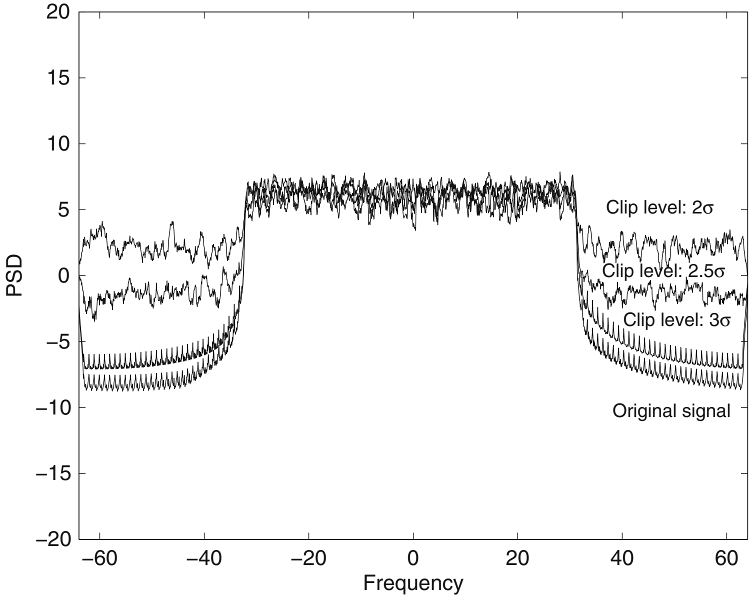


Fig. 4.54 Spectrum

$$S(f) = \frac{4\sqrt{6} f_o^3 e^{-l^2/2}}{9\pi^2 f^4} \tag{4.77}$$

for $f > f_0$, where we can see that if clip level is low, the spectral regrowth is high. Also, spectral regrowth decays with the frequency. This approach can be extended for impulsive nature, since bursty nature of clipping should be taken into account to get a realistic measure because instantaneous power spill may be averaged over and mislead system designers.

To remedy the out-of-band distortion, peak windowing may be applied, which multiplies large signal peak with a Gaussian shaped window. In practice, it convolves the original OFDM spectrum irrespective of number of subcarriers with the spectrum of the applied window. Cosine, Kaiser, and Hamming windows are suitable windows that have narrowband in frequency and not long in time domain. Peak windowing brings large reduction of PAPR regardless of number of subcarriers and has no effect to coding rate. As width of the window increases, the spectral regrowth decreases. BER performance on the other hand degrades with peak windowing, since it also distorts larger part of the unclipped signal.

This nonlinearity introduced to the system makes the system more vulnerable to errors. Hence, coding and scrambling techniques can be utilized along with clipping to increase the bit error rate performance and spectral efficiency.

4.7.3 Other Methods

Numerous techniques have been proposed to solve the PAPR problem. PAPR reduction techniques may fall into three categories: clipping, block coding, or peak cancelation where peak cancelation is an iterative way to subtract a known signal from the original signal to cancel the peaks. Block coding on the other hand adds redundancy to produce codes with low peaks. We review such existing techniques below.

- Block coding is one method to produce codes that achieves low PAPR. Of course, lower PAPR also limits the achievable code rate. Golay complementary sequences are a structured way to produce low PAPR codes with good FEC capabilities. Additional subcarriers are utilized in coding in order to achieve the error correction and PAPR reduction without distortion of the signal. If subcarriers have large amplitudes on the other hand, spectrum efficiency is poor and it requires large generation matrix.
- Tone reservation (TR) is one variant of block coding technique in which additional subcarriers carry no data but reserved for PAPR reduction. An effective cancelation signal in time domain can be found from only some number of additional subcarriers (aka reserved tones or null subcarriers) in the frequency domain

$$\hat{x}[n] = x[n] + c[n] = \mathrm{IFFT}(X_k + C_k), \tag{4.78}$$

where C_k stands for additional subcarriers. The new PAPR is

$$\text{PAPR}(\hat{\mathbf{x}}) = \frac{||\mathbf{x}+\mathbf{c}||_\infty^2}{E(||\mathbf{x}||_2^2)/N}, \tag{4.79}$$

since receiver ignores these additional subcarriers to recover the data. And one can see that $c[n]$ can be optimized to reduce the PAPR. Also PAPR reduction performance increases as the number of additional subcarriers increases, since the probability of constructing a cancelation signal increases. Notice also that distortion occurs only in the additional subcarriers, but not in the data carrying subcarriers. To construct the cancelation signal, TR employs a signal design algorithm, since signal must be designed in the frequency domain for its effect in the time domain. There are trial and error processes or computationally complex optimization procedures. The transmitter first checks for the peaks then for each peak TR method is performed. After peak cancelation, the composite signal is re-checked for secondary peaks that may appear during the peak cancelation.

- Selective mapping (SLM) utilizes redundant information and generates multiple instances of the same OFDM symbol to select the one with minimum PAPR. SLM reduces the PAPR by 2–3 dB, but requires side information at the receiver as well as multiple IFFT operation. The transmission of the side information is used to indicate the masking pattern, since the OFDM symbol is multiplied by K ($K > 1$) different available phase vectors that has N elements, each corresponds to each of the N subcarriers. This generates K statistically dependent OFDM symbols and the selected symbol is commonly referred to as the selected phase vector.
- Partial transmit sequence (PTS) creates M subblocks, each of them are subjected to L-point IFFT and then multiplied by a phase vector to minimize the PAPR. At the end, subblocks are summed and transmitted. This way optimum phase can be created per subblock, but search complexity is exponential with the number of blocks and side information is required.
- Dynamic range increase is another PAPR mitigation method proposed to WiMAX-m (IEEE 802.16m) that increases the PA's dynamic range to overcome high PAPR with low complexity. This is performed by envelope tracking so that VCC to the PA is raised to accommodate large peaks in the linear region.
- Active constellation extension (ACE) reduces the peak power by changing the signal constellation without affecting the BER performance, since minimum distance is preserved. For example in QPSK, there are four possible constellation points for each subcarrier where the transmitted bits could be mapped as seen in Fig. 4.55. These four points lie in each quadrant in the complex plane and are equidistant from the axis. A received data is assigned according to the quadrant in which the symbol is observed. Errors only occur if the received sample is mapped to one of the other three quadrants. Modification of the constellation points within the quarter plane is allowed in ACE, since this adds additional sinusoidal signals at the particular frequency to the transmitted signal. With correct

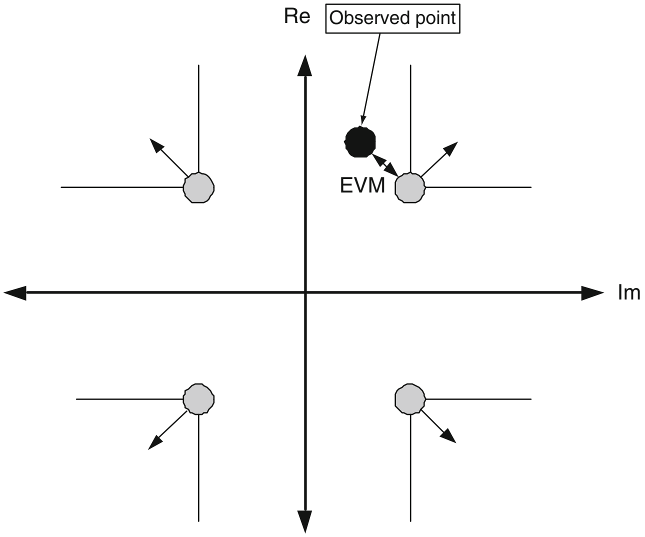


Fig. 4.55 ACE

adjustment, these signals are used to cancel time-domain peaks in the transmitted OFDM signal.

In WiMAX, PAs must deliver more power, be more linear, and have the ability to handle a PAPR around 10 dB. This brings tight EVM[19] requirement around -31 dB, based on 1% packet error rate. This enforces more linear component in the system and contributes WiMAX's longer range with stringent receiver noise figure (7 dB maximum).

4.8 Application: IEEE 802.11a

OFDM came to prominence with IEEE 802.11a/g wireless local area networking standard.[20]

IEEE 802.11a MAC is based on a random access scheme called CSMA/CA (carrier sense multiple access/collision avoidance[21]) protocol, which only grants the

[19] "The error vector magnitude or EVM is a measure used to quantify the deviation in the constellation points from the ideal locations due to various imperfections in the implementation such as carrier leakage, PA and D/A nonlinearity, phase noise, etc. These cause the actual constellation points to deviate from the ideal locations as seen in Fig. 4.55. Basically, EVM is a measure of how far the points are from the ideal locations."

[20] IEEE 802.11a-1999 (aka 802.11a) is an amendment to the IEEE 802.11 specification for operation in 5 GHz and IEEE 802.11g-2003 is another amendment to provide backward compatibility with IEEE 802.11b in 2.4 GHz.

[21] Note that classic Ethernet uses CSMA/CD – collision detection.

channel by contention. In CSMA/CA, a wireless node that wants to transmit performs the following:

Step 1 Listen the channel.
Step 2 Transmit if channel is idle.
Step 3 If channel is busy, wait until transmission stops plus a contention period, which is a random period to ensure fairness. Contention period is quantified with a back-off counter where a node decrements the back-off counter if it detects channel idle for a fixed amount of time.
Step 4 Node transmits when back-off counter is zero.
Step 5 If the transmission is unsuccessful – no ACK, contention window is selected from a random interval, which is twice the previous random interval. The process is repeated until it gets a free channel.

Hidden node problem is solved with request-to-send (RTS) and clear-to-send (CTS) message exchange before transmitting the actual packet. If a station receives CTS packet from destination after sending the RTS packet, it reserves the channel for the duration of its packet plus a ACK. Packet size is upper bounded with the maximum packet size. Hence, if packet size is small, channel utilization is very low.

OFDM PHY transmits MAC protocol data units (MPDUs) as directed by the MAC layer. The OFDM PHY is composed of two elements: the physical layer convergence protocol (PLCP) and the physical medium dependent (PMD) sublayers. The PLCP prepares PLCP frame from MPDUs for transmission. The PLCP also delivers received frames from the air medium to the MAC layer. The PMD provides modulation and demodulation of the frame transmissions.

PLCP layer frame is illustrated in Fig. 4.56 and related key parameters are depicted in Table 4.2. There are two preamble sequences, each are two symbol length. First preamble contains ten short training sequences (STSs) and second preamble has two long training sequences (LTSs). First preamble is mostly for signal detection, automatic gain control, diversity selection, timing acquisition, and coarse frequency acquisition. Second preamble is used for channel estimation and fine frequency acquisition. Each symbol has a guard interval and first symbol after the second preamble contains information about rate, length, tail, service, etc., and always coded with a BPSK with coding rate of 1/2. Within each symbol also there are pilot subcarriers for frequency offset estimation and timing as seen in Fig. 4.57.

Table 4.3 shows the achievable physical layer data rates for IEEE 802.11a/g with convolutional coding. These numbers are raw rates and typically net throughput

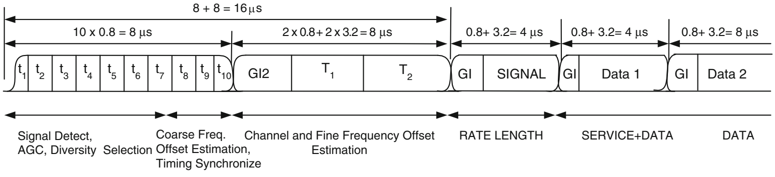


Fig. 4.56 Format of an OFDM frame (© IEEE)

Table 4.2 IEEE 802.11a parameters

IEEE 802.11a Parameters	Value
Number of data subcarriers N_{SD}:	48
Number of pilot subcarriers ND_{SP}:	4
Number of subcarriers, total N_{ST}:	52
Subcarrier frequency spacing Δ_F:	0.3125 MHz (=20 MHz/64)
IFFT/FFT period T_{FFT}:	3.2 µs ($1/\Delta_F$)
PLCP preamble duration T_{PREAMBLE}:	16 µs ($T_{\mathrm{SHORT}} + T_{\mathrm{LONG}}$)
Duration of the SIGNAL T_{SIGNAL}:	4.0 µs ($T_{\mathrm{GI}} + T_{\mathrm{FFT}}$)
GI duration T_{GI}:	0.8 µs ($T_{\mathrm{FFT}}/4$)
Training symbol GI duration T_{GI2}:	1.6 µs ($T_{\mathrm{FFT}}/2$)
Symbol interval T_{SYM}:	4 µs ($T_{\mathrm{GI}} + T_{\mathrm{FFT}}$)
Short training sequence duration T_{SHORT}:	8 µs ($10 \times T_{\mathrm{FFT}}/4$)
Long training sequence duration T_{LONG}:	8 µs ($T_{GI2} + 2 \times T_{\mathrm{FFT}}$)
Signal Bandwidth W:	16.66 MHz

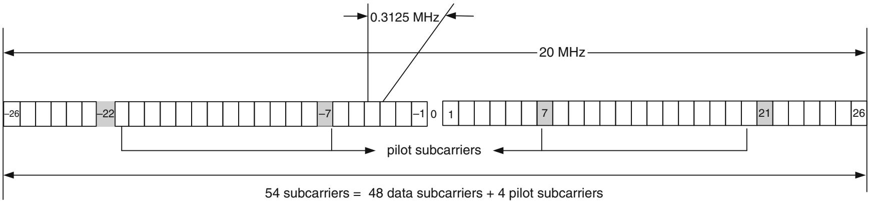


Fig. 4.57 OFDM subcarrier allocation for data and pilot

Table 4.3 Achievable physical layer data rates with IEEE 802.11a

Mode	Modulation	Code rate	Data rate (Mbps)
1	BPSK	1/2	6
2	BPSK	3/4	9
3	QPSK	1/2	12
4	QPSK	3/4	18
5	16QAM	1/2	24
6	16QAM	3/4	36
7	64QAM	2/3	48
8	64QAM	3/4	54

is around 28 Mbps for 54 Mbps (with 54% inefficiency), which is achieved with 64QAM modulation and 3/4 coding rate in 20-MHz bandwidth.

4.9 Summary

In this chapter, we give principles of OFDM. We discuss OFDM theory and key components of OFDM transmission: coding, synchronization, channel estimation, equalization, and peak-to-average-power ratio. More detailed information about

OFDM and its applications can be found in *Multicarrier Digital Communications: Theory and Applications of OFDM*, published by Springer in 2004. Highlights about OFDM as a summary are:

- OFDM creates orthogonal spectral efficient low rate carriers in order to transmit high-rate signals.
- OFDM utilizes cyclic prefix in the guard interval in order to guarantee no ISI and ICI.
- OFDM is insensitive to timing offset but sensitive to frequency offset and phase noise.
- OFDM utilizes known preambles or pilot symbols for coherent detection and synchronization.
- OFDM does not need a time domain equalizer and needs only a simple frequency domain equalizer to correct amplitude and phase changes.
- OFDM may utilize a time domain equalizer to shorten the guard period.
- OFDM systems can utilize several coding schemes: Reed Solomon coding, convolutional coding, concatenated coding, Trellis coding, turbo coding, and LDPC coding.
- OFDM may show uncontrolled high peaks. Peak-to-average power ratio is a major problem and one way to alleviate is Clipping the high peaks: Clipping is a nonlinear process and it introduces distortion both inside and outside the given signal bandwidth.

References

1. Bahai, A., Saltzberg, B., Ergen, M., *Multi-Carrier Digital Communications: Theory and Publications of OFDM,* Springer, New york, 2004.
2. Goldsmith, A., *Wireless Communications,* Cambridge University Press, Cambridge, 2005.
3. Hanzo, L, Webb, W., Keller, T., *Single- and Multi-carrier Quadrature Amplitude Modulation,* Wiley, New York, 2000.
4. Van Nee, R., Prasad, R., *OFDM for Wireless Multimedia Communications,* Artech House, Boston, 2000.
5. Jha, U. S., Prasad, R., *OFDM Towards Fixed and Mobile Broadband Wireless Access,* Artech House, Boston, 2007.
6. Ahn, J., Lee, H. S., "Frequency domain equalization of OFDM signal over frequency non selective Rayleigh fading channels," *Electronic Letters,* pp. 476–1477, 1993.
7. Armstrong, J., "Analysis of new and existing methods of inter-carrier interference due to carrier frequency offset in OFDM," *IEEE Transactions on Communication,* vol. 47, no. 3, pp. 365–369, 1999.
8. Baml, R. W., Fischer, R. F. H., Huber, J. B., "Reducing the peak to average power ratio of multicarrier modulation by selected mapping," *IEE Electronics Letters,* vol. 32, no. 22, pp. 2056–2057, 1997.
9. Bhatti, S. N., "WiFi Lecture notes for M.Sc. data communication networks and distributed systems D51 – Basic communications and networks," Department of Computer Science, University College, London, 1994.
10. Chang R. W., "Synthesis of band limited orthogonal signals for multichannel data transmission," *Bell System Technical Journal*, vol. 45, pp. 1775–1796, 1996.

11. Salzberg, B. R., “Performance of an efficient parallel data transmission system,” *IEEE Transactions on Communications*, vol. 15, pp. 805–813, 1967.
12. Mosier, R. R., Clabaugh, R.G., “A Bandwidth efficient binary transmission system,” *IEEE Transactions*, vol. 76, pp. 723–728, 1958.
13. MacWilliams, F. J., Sloane, N. J. A., *The Theory of Error Correcting Codes*, North-Holland, New York, 1977.
14. Firmanto, W. T., Gulliver, T. A., “S code combining of Reed-Muller codes in an indoor wireless environment,” *Wireless Personal Communications*, vol. 6, pp. 359–371, 1997.
15. Haccoun D., Begin G., “High-rate punctured convolutional codes for Viterbi and sequential decoding,” *IEEE Transactions on Communication*, vol. 37, pp. 1113–1125, 1989.
16. Tipler P., *Physics for Scientists and Engineers*, 3rd edn., Worth Publishers, New York; pp. 464–468, 1991.
17. Rappaport, T. S., *Wireless Communications Principles and Practice*, IEEE Press, New York, pp. 169–177, 1996.
18. Prasad R., Van Nee, R., *Synthesis OFDM for Wireless Multimedia Communications*, Artech House, Boston, pp. 80–81, 2000.
19. Cox, D. C., Schmidl, T. M., “Robust frequency and timing synchronization on OFDM,” *IEEE Transactions on Communications*, vol. 45, no. 12, pp. 1613–1621, 1997.
20. Classen, F., Meyr, H., Sehier, P., “Maximum likelihood open-loop carrier synchronizer for digital radio,” *Proceedings of ICC*, pp. 493–497, 1993.
21. Classen, F., Meyr, H., “Frequency synchronization algorithms for OFDM systems suitable for communications over frequency selective fading channels,” *Proceedings of IEEE Vehicular Technology Conference (VTC)*, pp.1655–1659, 1994.
22. Proakis, J. G., *Digital Communications*, 3rd edn., Prentice Hall, Englewood Cliffs, New Jersey, 1995.
23. Cavers, J. K., “An analysis of pilot symbol assisted modulation for Rayleigh fading channels,” *IEEE Transactions on Vehicular Technologies*, vol. 40, no. 4, pp. 686–693, 1991.
24. Maseng, T., Tufvesson, F., “Robust Pilot assisted channel estimation for OFDM in mobile cellular systems,” *IEEE 47th Vehicular Technology Conference Technology in Motion*, vol. 3, pp. 1639–1643, 1997.
25. Cimini, J., Sollenberger, N. R., “OFDM with diversity and coding for advanced cellular Internet services,” *IEEE Conference Proceedings VTC*, 1998.
26. Cimini, L. J., “Analysis and simulation of digital mobile channel using orthogonal frequency division multiplexing,” *IEEE Transactions on Communication*, vol. 33, no. 7, pp. 665–675, 1985.
27. Eetvelt, P. V., Wade, G., and Tomlinson, M., “Peak to average power reduction for OFDM schemes by selective scrambling,” *IEE Electronics Letters*, 1996.
28. Fernando, W. A. C., Rajatheva, R. M. A. P., “Performance of COFDM for LEO satellite channels in global mobile communications,” *IEEE Conference Proceedings VT*, vol. 1, pp. 412–416, 1998.
29. Gudmundson, M., Anderson, P. O., “Adjacent channel interference in an OFDM system,” *IEEE Conference Proceedings VTC*, 1996.
30. Jayalath, A. D. S, “Application of orthogonal frequency division multiplexing with concatenated coding in wireless ATM,” Master Thesis, Asian Institute of Technology, Bangkok, Thailand, 1997.
31. Barton, S. K., Jones, A. E., Wilkinson, T. A., “Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes,” *Electronics Letters*, vol. 25, no. 25, pp. 2098–2099, 1994.
32. Kafle, P., “Performance of parallel concatenated interleaved codes in correlated multipath fading channels,” Master Thesis, Asian Institute of Technology, Bangkok, Thailand, 1998.
33. Baum, K. L., “A synchronous coherent OFDM air interface concept for high data rate cellular systems,” *IEEE Conference Proceedings VTC*, pp. 2222–2226, 1998.
34. Li, R., Stette, G., “Time limited orthogonal multicarrier modulation schemes,” *IEEE Transactions on Communications*, vol. 4, pp. 1269–1272, 1995.

35. Li, X., Cimini, L. J., "Effects of Clipping and Filtering on the Performance of OFDM," *Communication Letters*, vol. 2, no. 5, pp. 131–133, 1998.
36. Li, X., Ritcey, J. A., "M-sequences for OFDM PAPR reduction and error correction," *Electronic Letters*, vol. 33, pp. 545–546, 1997.
37. May, T., Rohling, H., "Reducing the peak to average power ratio in OFDM radio transmission systems," *IEEE Conference Proceedings VTC*, pp. 2474–2478, 1998.
38. Moose, P. H., "A technique for orthogonal frequency division multiplexing frequency offset correction," *IEEE Transactions on Communication*, vol. 42, no. 10, pp. 2908–2914, 1994.
39. Muller, S. H., Huber, J. B., "OFDM with reduced peak to average power ratio by optimum combination of partial transmit sequences," *IEE Electronics Letters*, vol. 33, no. 5, pp. 368–369, 1997.
40. Muller, S. H., Huber, J. B, "OFDM with reduced peak to average power ratio by optimum combination of partial transmit sequences," *IEE Electronics Letters*, vol. 33, no. 5, pp. 368–369, 1997.
41. Muschallik, C., "Improving OFDM reception using an adaptive Nyquist windowing," *IEEE Transactions on Consumer Electronics*, vol. 42, no. 3, 1996.
42. Bossert, M., Donder, A., Nogueroles, R., Zyablov, V., "Improved performance of a random OFDMA mobile communication system," *IEEE Conference Proceedings VTC*, 1998.
43. Lopes, L. B., O'Neill, R., "Envelope variations and spectral splatter in clipped multicarrier signals," *IEEE Conference Proceedings PMIRC*, pp. 71–76, 1995.
44. Kuchenbecker, H. P., Pauli, M., "On the reduction of the out of band radiation of OFDM signals," *IEEE Conference Proceedings ICC*, vol. 3, pp. 1304–1308, 1998.
45. Moeneclaey, M., Pollet, T., Van Bladel, M., "BER sensitivity of OFDM systems to carrier frequency offset and wiener phase noise," *IEEE Transactions on Communications*, vol. 43, no. 2, pp. 191–193, 1995.
46. Grunheid, R., Rohling, H., "Performance of an OFDM-TDMA mobile communication system," *IEEE Conference Proceedings VTC*, pp. 1589–1593, 1996.
47. Shelswell, P., "The COFDM modulation system: The heart of digital audio broadcasting," *Electronics & Communication Engineering Journal*, pp. 127–135, 1995.
48. Barton, S., Orriss, J., Shepherd, S., "Asymptotic limits in peak envelope power reduction by redundant coding in orthogonal frequency division multiplexing modulation," *IEEE Transactions on Communication*, vol. 46, no. 1, pp. 5–10, 1998.
49. Shrestha, N., "Compensation of co-channel interference in equalization," Master's thesis, Asian Institute of Technology, 1995.
50. Cioffi, J. M., Tellado, J., "Peak power reduction for multicarrier transmission," *IEEE Conference Proceedings Globecom*, 1998.
51. Van Nee, R., Wild, A., "Reducing the peak to average power ratio of OFDM," *IEEE Conference Proceedings VTC*, pp. 2072–2076, 1998.
52. Goldfeld, L., Wulich, D., "Reduction of peak factor in orthogonal multicarrier modulation by amplitude limiting and coding," *IEEE Transactions on Communication*, vol. 47, no. 1, pp. 18–21, 1999.
53. Forney, G. D., "Convolutional codes II: Maximum-likelihood decoding," *Information Control*, vol. 25, pp. 222–226, 1974.
54. Gilhousen, K. S., et. al., "Coding systems study for high data rate telemetry links," Final Contract Report, N71-27786, Contract No. NAS2-6024, Linkabit Corporation, La Jolla, CA, 1971.
55. Heller, J. A., Jacobs, I. M., "Viterbi decoding for satellite and space communications," *IEEE Transactions on Communication*, vol. 19, pp. 835–848, 1971.
56. Larsen, K. J., "Short convolutional codes with maximal free distance for rates 1/2, 1/3, and 1/4," *IEEE Transactions on Information Theory*, vol. 19, pp. 371–372, 1973.
57. Odenwalder, J. P., *Optimum decoding of convolutional codes*, PhD dissertation, Department of Systems Sciences, School of Engineering and Applied Sciences, University of California at Los Angeles, 1970.
58. Viterbi, A. J., "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, pp. 260–269, 1967.

59. Gumas, C., “Turbo codes rev up error-correcting performance (Part 1),” *PE&IN*, pp. 61–66, 1998.
60. Gumas, C., “Turbo codes build on classic error-correcting codes and boost performance (Part 2),” *PE&IN*, pp. 54–63, 1998.
61. Gumas, C., “Turbo codes propel new concepts for superior codes (Part 3),” *PE&IN*, pp. 65–70, 1998.
62. Gumas, C., “Win, place, or show, turbo codes enter the race for next generation error-correcting systems (Part 4),” *PE&IN*, pp. 54–62, 1998.
63. Borjesson, P. O., Sandell, M., Van de Beek, J. J., “ML estimation of time and frequency offsets in OFDM systems,” IEEE Transactions on Signal Processing, vol. 45, no. 7, pp. 1800–1805, 1997.
64. Bingham, J. A. C., “Multicarrier modulation for data transmission: An idea whose time has come,” *IEEE Communications Magazine*, vol. 28, no. 5, pp. 5–14, 1990.
65. Cavers, J. K., “An analysis of pilot-symbol assisted modulation for Rayleigh-fading channels,” *IEEE Transactions on Vehicular Technology*, vol. 40, no. 4, pp. 686–693, 1991.
66. Chang, R. W., “Synthesis of band-limited orthogonal signals for multichannel data transmission,” *Bell System Technical Journal*, vol. 45, pp. 1775–1796, 1966.
67. Gibby, R. A., Chang, R. W., “Theoretical study of performance of an orthogonal multiplexing data transmission scheme,” *IEEE Transactions on Communication* vol. 16, no. 4, pp. 529–540, 1968.
68. Chini, A., *“Multicarrier modulation in frequency selective fading channels,”* PhD thesis, Carleton University, 1994.
69. Classen, F., Meyr, H., “Frequency synchronization algorithms for OFDM systems suitable for communication over frequency-selective fading channels,” *Proceedings of the IEEE VTC*, pp. 1655–1659, 1994.
70. Adami, O., Daffara, F., “A new frequency detector for orthogonal multicarrier transmission techniques,” *Proceedings of the IEEE VTC*, pp. 804–809, 1995.
71. Borjesson, P. O., Edfors, O., Sandell, M., Wilson, S. K., Van de Beek, J. J., “OFDM Channel Estimation by singular value decomposition,” *IEEE Transactions on Communication*, vol. 46, no. 7, pp. 931–939, 1998.
72. Borjesson, P. O., Edfors, O., Sandell, M., Wilson, S. K., Van de Beek, J. J., “Analysis of DFT-based channel estimators for OFDM,” *Wireless Personal Communications, Kluwer, Dordrecht*, 1998.
73. Engels, V., Rohling, H., “Multilevel differential modulation techniques (64-DAPSK) for multicarrier transmission systems,” *European Transactions on Telecommunications*, vol. 6, no. 6, pp. 633–640, 1995.
74. Calvo, M., Garcia Armada, A., “Phase noise and sub-carrier spacing effects on the performance of an OFDM communication system,” *IEEE Communications Letters*, vol. 2, no. 1, pp. 11–13, 1998.
75. Veeneman, D., Gross, R., “Clipping distortion in DMT ADSL systems,” *Electronics Letters*, vol. 29, no. 24, pp. 2080–2081, 1993.
76. Barton, S. K., Jones, J., Wilkinson, T. A., “Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes,” *Electronics Letters*, vol. 30, no. 25, pp. 2098–2099, 1994.
77. Lodge, J. H., Moher, M. L., “TCMP – A modulation and coding strategy for Rician-fading channels,” *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 9, pp. 1347–1355, 1989.
78. Huber, J. B., Mller, S. H., “A comparison of peak power reduction schemes for OFDM,” *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM’97)*, pp. 1–5, 1997.
79. Saltzberg, B. R., “Performance of an efficient parallel data transmission system,” *IEEE Transactions on Communication*, vol. 15, no. 6, pp. 805–811, 1967.
80. Couch, L. W., *Digital and Analog Communication Systems*, 4th ed. Macmillan, New York, 1993.

81. Haykin, S., *Communication Systems*, 3rd ed., Wiley, New York, 1994.
82. Proakis, J. G., *Digital Communications*, 3rd ed., WCB/McGraw-Hill, Boston, MA, 1995.
83. Berrou, C., "Some clinical aspects of turbo codes," *International Symposium on Turbo Codes*, pp. 26–31, 1997.
84. Benedetto, S., Montorsi, G., "Design of parallel concatenated convolutional codes," *IEEE Transactions on Communication*, vol. 44, 1996.
85. Benedetto, S., Divsalar, D., Montorsi, G., Pollara, F., "Communication algorithm for continuous decoding of turbo codes," *Electronic Letters*, vol. 32, no. 4, 1996.
86. Pyndiah, R. M., "Near-optimum decoding of product codes: block turbo codes," *IEEE Transactions on Communication*, vol. 46, pp. 1003–1010, 1998.
87. Foschin, G. J., "Turbo Layered space-time architecture for wireless communication in a fading environment when using multiple antennas," *Bell Labs Technical Journal*, vol. 1, no. 2, pp. 41–59, 1996.
88. Reed Solomon Codes. http://komodo-industries.com/basic_reed-solomon_tutorial.html.
89. Leiner, B. M. J., LDPC Codes – a brief tutorial. http://users.tkk.fi/pat/coding/essays/ldpc.pdf.
90. Ergen, M., Varaiya, P., "Throughput analysis and admission control in IEEE 802.11a," *Springer Mobile Networks and Applications*, vol. 10, no. 5., pp. 705–706, October 2005.
91. Cripps, S. C., *RF Power Amplifiers for Wireless Communications*, Artech House, Boston, 1999.

Chapter 5
Principles of OFDMA

5.1 Overview

Let us give an overview how wireless technology takes a smooth coarse from single-carrier to OFDMA to understand the necessity for robust communication and better efficiency.

To date, single-carrier modulation was the prominent transmission scheme for cellular networks. It is important to know how single-carrier is susceptible to wireless multipath channel distortion. Multipath may cause delay in some paths and arrival may coincide with successor signals. This is called delay spread in the literature and overlay of delayed signal onto current is known as *intersymbol-interference*.

In single-carrier systems, as data rate increases symbol time decreases from Nyquist-Shannon theorem; significant delay caused by multipath on the order of few symbol times may spill the delayed symbol into later symbols. Each multipath arrives with an amplitude change and phase shift. And when combined in the receiver, some frequencies within the signal experience constructive and destructive interference. These induce increase in the error rate.

Single-carrier systems compensate for the channel distortion via time equalization. Typical time equalization considers a channel inversion by determining the channel response with transmitted known pilot sequence. CDMA systems on the other hand utilize rake receivers to resolve the individual paths and then align them in time to enhance signal quality.

Complexity of equalizer in single-carrier system increases with data rates, since as data rates increase, the receiver needs to get more frequent samples to compensate for the delay spread and consequently sample clock (t) decreases. These increase the number of delay taps (see Fig. 5.1) in the equalizer and makes it almost impossible to meet rates above 100 Mbps.

OFDM on the other hand converts a single spectrum into many narrower subcarriers and transmits the data in parallel streams. As a result, subcarrier data rate is lower than the actual targeted data rate, which makes symbol length longer and reduces the complexity in the equalizer where there is an equalizer for each subcarrier.

DOI: 10.1007/978-0-387-68192-4_5, © Springer Science+Business Media LLC 2009

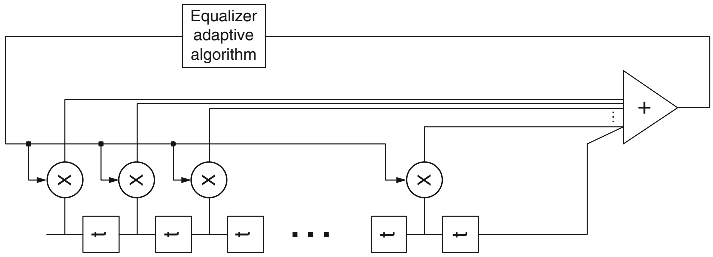


Fig. 5.1 Time domain channel equalizer

Parallel and tightly spaced subcarriers are achieved by fast fourier transform (FFT). ISI is eliminated by cylic prefix (CP) which precedes the OFDM symbol. If CP length is sufficient, preceding symbols do not leak into the actual symbol.

The complexity comparison between OFDM and single-carrier transmission considers complexity between FFT and single-carrier time domain equalizer. For instance, complexity of 64 point radix-4 FFT in IEEE 802.11a OFDM is 96 million multiplications per second, but 16 taps OQPSK or GMSK Equalizer for same data rates above needs 768 million multiplications per second. Notice that this structure is not confined in terms of achieving higher data rates. The OFDM benefits comes with three major drawbacks: frequency offset, PAPR, and CP overhead. These are detailed in the previous chapter.

OFDMA is the multiplexing scheme for OFDM. OFDMA inherits all the advantages and disadvantages of OFDM and exhibits some new features. Multiplexing provides packing many user packets into one frame for uplink and downlink. As a result, multiplexing scheme becomes very efficient in the sense that overheads caused by interframe spacing is minimized.

5.1.1 Random Access: CSMA-OFDM

For instance, in IEEE 802.11 Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) scheme is used as the multiplexing method. CSMA employs listen-before-talk scheme and whoever gets the transmission right only speaks or listens for itself. Users backoff for a random time and then transmit. If they detect collision they backoff more and transmit again. Of course, if they detect transmission activity in the channel, they either abandon transmitting or stop decrementing their backoff counter. As a result, there is a random waiting time between successful transmissions.

In CSMA/CA, if a user transmits, it captures the channel for itself, capture duration is determined by the packet size and modulation scheme. Typically, selected modulation scheme depends on the channel condition, but the packet size is user specific. Look at Table 5.1 to see the distribution of packet size over the Internet. In downlink, more than 50% of packets are less than 1,000 bytes in size, and in uplink,

Table 5.1 Packet size distribution in Internet

Bytes	Downstream %	Upstream %
0–64	14.68	58.49
65–127	13.87	29.73
128–255	7.25	1.72
256–511	6.44	3.98
512–1023	13.59	3.37
1024–1518	44.17	2.70

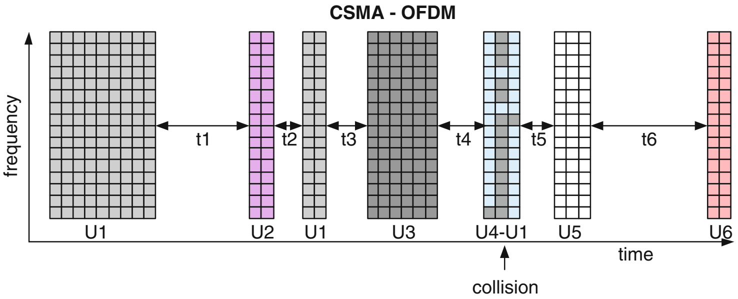


Fig. 5.2 CSMA-OFDM: There are 6 users (U), and CSMA scheme has random time intervals between frames (ti) and random packet sizes

more than 97% of packets are less than 1,000 bytes. As a result, overhead increases if a user transmits a packet that is smaller in size. Not surprisingly, real-world efficiency of IEEE 802.11a systems is approximately 50%, which means that typical throughput is about 25–30 Mbps for a data rate of 54 Mbps (see Fig. 5.2).

5.1.2 Time Division: TDMA-OFDM

On the other hand in TDMA-OFDM, the waiting time between each frame is fixed and minimal. Also, there is no contention for transmission since each user is assigned within a frame by a centrally coordinated mechanism residing in the base station called scheduler.

TDMA-OFDM divides the time into orthogonal transmission opportunity for each user as seen in Fig. 5.3. Each user transmits with OFDM using all the subcarriers in allocated fixed timeslots. When compared with CSMA-OFDM, it is efficient, since TDMA-OFDM avoids contention and collisions with prefixed and controlled transmission opportunity.

On the other hand, fixed transmission opportunity fixes the packet size with available modulation. TDMA-OFDM systems are suitable for constant bit rate systems. But, a general packet-based system requires more flexibility in terms of transmission opportunity, since the packet size is variable. Modulation is coarse and

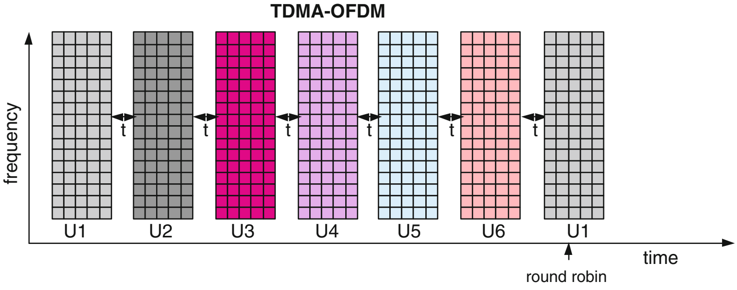


Fig. 5.3 TDMA-OFDM: There are 6 users (U), and TDMA scheme has fixed time intervals between frames (t) and fixed packet sizes

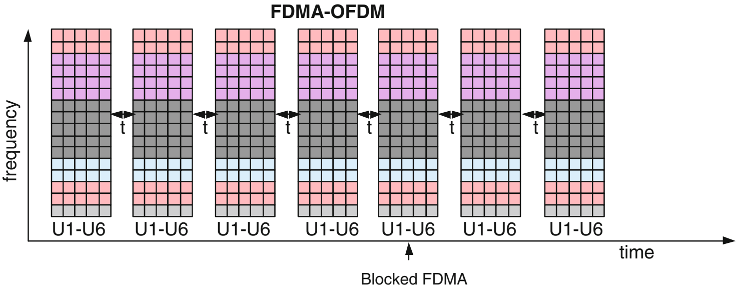


Fig. 5.4 FDMA-OFDM: There are 6 users (U), and Block-FDMA scheme has fixed time intervals between frames (t) and fixed subcarrier allocation

applied to all subcarriers. If the channel is low quality for a user in the time slot, either transmission experiences high bit error rate with the higher modulation, or lower modulation is selected by reducing the bit rate. Also, if there is no data to transmit, the bandwidth is wasted.

5.1.3 Frequency Division: FDMA-OFDM

FDMA-OFDM is similar to TDMA-OFDM. FDMA-OFDM creates orthogonal resource by dividing the available subcarriers into fixed sets where each set is used by a user as seen in Fig. 5.4. Fixed allocation does not change over time but the number of subcarriers associated to a user may differ. Albeit, allocation is fixed, which prevents leveraging the multiuser diversity.

There are two types of allocation: localized FDMA (LFDMA) or interleaved FDMA (IFDMA). Block FDMA allocates sets to adjacent subcarriers. Interleaved FDMA, on the other hand, interleaves the subcarrier when defining the set.

5.1.4 Code Division: MC-CDMA

CDMA with OFDM waveform is another way of implementing multiplexing. As in traditional CDMA, signal is spread over frequency or time domain. CDMA with OFDM waveform (aka multicarrier CDMA) spread the symbol over time slots or subcarriers. Figure 5.5 shows a time spreading representation of MC-CDMA. A data symbol is multiplied by a code and spread is established in the same subcarrier over time.

5.1.5 Space Division: SDMA-OFDM

SDMA system applies space division when creating the orthogonal resources. Each user experiences a different channel, and user-specific spatial signature is created by the channel and acts like spreading code in a CDMA system. Multiuser detection techniques known from CDMA can be applied in SDMA-OFDM. Figure 5.6 shows channels for SDMA-OFDM system. There is a user-specific spatial signature to identify the user in the receiver.

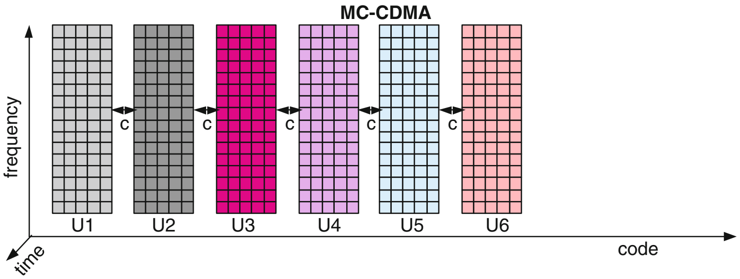


Fig. 5.5 MC-CDMA: There are six users (U), and time-spread MC-CDMA scheme has distance between codes (c) for orthogonality

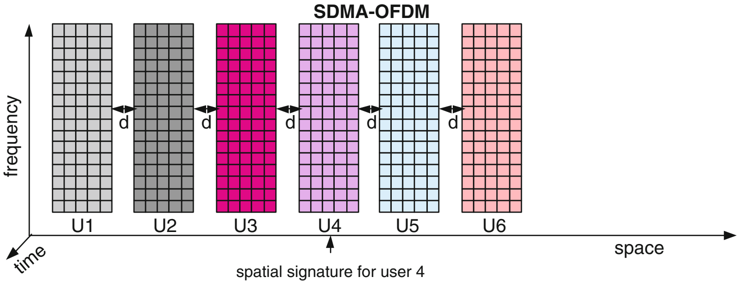


Fig. 5.6 SDMA-OFDM: There are 6 users (U), and SDMA scheme has physical distance between receivers (d) for orthogonality

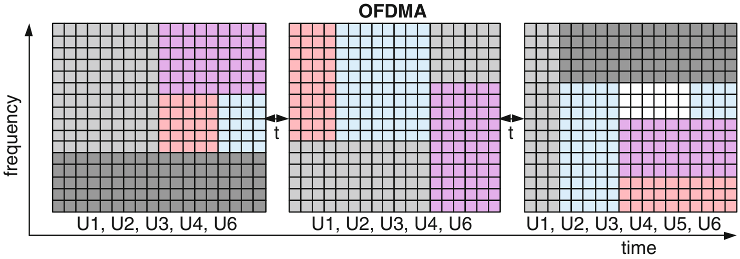


Fig. 5.7 OFDMA: There are 6 users (U), and OFDMA scheme has fixed distance between frames (t) and flexible slot and subcarrier allocation

5.1.6 OFDMA

OFDMA inherits the advantages of TDMA and FDMA schemes. It also provides flexibility as in CSMA in terms of packet sizes. User is scheduled dynamically only if it has packet to transmit according to how much it needs to transmit.

In OFDMA, on the other hand, a transmission is packed as seen in Fig. 5.7. Frame size is fixed as in TDMA and FDMA, but frame is shared by users. As a result, efficiency is increased considerably and air link efficiency becomes isolated from the traffic pattern of users. Sophisticated QoS scheduling can be applied with respect to widely varying applications, data rates, etc.

OFDMA scheduling may exploit diversity to increase the capacity of the air link. For instance, each user experiences different channel, as a result, a slot (subcarrier × time symbol) that is low in quality might not be low for others. OFDMA may exploit multiuser diversity by assigning the subcarriers according to channel quality of each user if channel side information is present in the transmitter. Also, modulation and power control may be tuned according to each subcarrier with respect to channel condition.

OFDMA scheduling may also be based on frequency diversity. An orthogonal code may be defined for each user to determine their hopping pattern across time and frequency. This way, fading of a user may be averaged over without any channel information. Also, in a typical cell deployment, same frequency is used in multiple spatially located cells. OFDMA may reduce the interference coming from the adjacent cells operating with the same frequency by assigning different sets of orthogonal codes in each cell.

This may lead to utilize OFDMA to deploy a single frequency network in several ways. In one way, the same frequency is used by the base stations but slot allocation is done with respect to a code that is user specific. Codes need not to be strictly orthogonal, which would otherwise result in low capacity system, since users are separated in distance, which makes the probability of interference minimal. In another way, some slots in a frame can be restricted to be used for the users that are not in the overlapping area of cells. As a result, the remaining slots are used to create orthogonal resources (typically three orthogonal resource is ample), and each adjacent cell uses one of the orthogonal resource in order to schedule its users, which resides in the overlapping region.

OFDMA system performance can be enhanced by exploiting space diversity as well. SDMA or MIMO support for OFDMA improves the capacity and performance significantly.

It is apparent that scheduling plays an important role in OFDMA systems. In this chapter we review a few scheduling algorithms that exploit either multiuser diversity or frequency diversity. We first assume a single cell with no adjacent cell interference, then we elevate the analysis to cover multicell with adjacent cell interference.

5.2 Multiuser Diversity and AMC

In a multiuser environment, a good resource allocation scheme leverages multiuser diversity and channel fading. The optimal solution for an OFDM system is to schedule the user with the best channel at each time. Although in OFDM case, the entire bandwidth is used by the scheduled user, this idea can also be applied to OFDMA system. Here, the channel is shared by the users, each owing a mutually disjoint set of subcarriers, via scheduling the subcarrier to a user with the best channel among others. Of course, the procedure is not simple since the best subcarrier of the user may also be the best subcarrier of another user who may not have any other good subcarriers. The overall strategy is to use the peaks of the channel resulting from channel fading. Unlike in the traditional view where the channel fading is considered to be an impairment, here it acts as a channel randomizer and increases multiuser diversity.

Adaptive modulation and coding (AMC) is utilized to take advantage in the randomness of the channel. When the channel is in good condition, the transmission is performed with higher data rates, and when the channel is poor, the transmission rate is lowered with small constellation and low-rate codes. The channel side information is feeded to transmitter in order to control transmit power, transmit constellation, and the coding rate.

For example, let us say data rate and bit error rate requirements of users are as follows

User	1	2	3	4
Rate (bits)	12	6	6	8
BER	1e-2	1e-2	1e-4	1e-4

and the channel for a given time instant is noted as in Fig. 5.8. Optimal resource allocation and adaptive modulation would give a distribution as in Fig. 5.9.

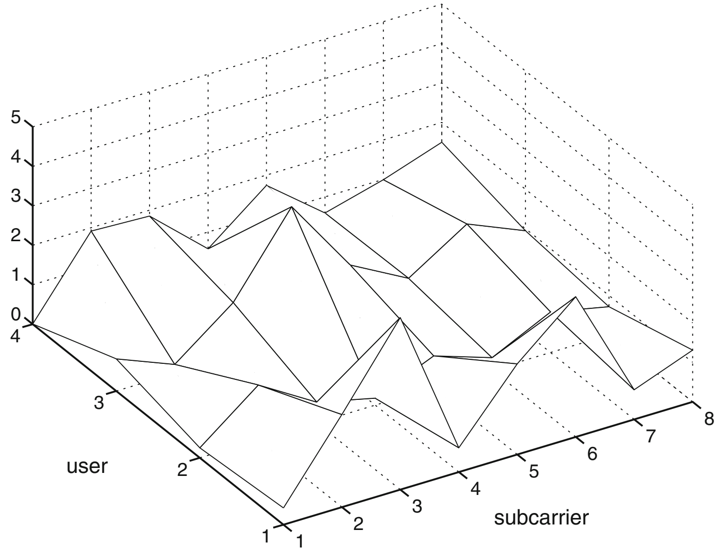
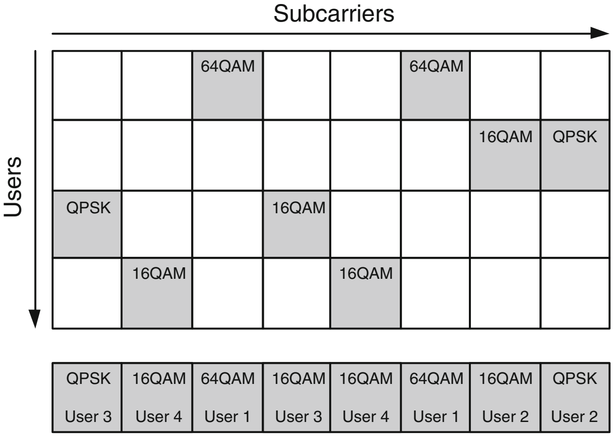


Fig. 5.8 A time instance of wireless channel for each user



		64QAM			64QAM		
						16QAM	QPSK
QPSK			16QAM				
	16QAM			16QAM			

QPSK User 3	16QAM User 4	64QAM User 1	16QAM User 3	16QAM User 4	64QAM User 1	16QAM User 2	QPSK User 2

Fig. 5.9 Optimal resource allocation and bit loading

5.3 OFDMA System Model and Formulation

Now, we introduce the OFDMA system. The scalable OFDMA introduced below is a new technique introduced in WiMAX to host different channel bandwidths with fixed subcarrier spacing.

5.3.1 Scalable OFDMA

The scalable OFDMA (SOFDMA)[1] enables standard-based solutions to deliver optimum performance in channel bandwidths ranging from 1.25 to 20 MHz with fixed subcarrier spacing for mobile environment. Mobility affects Doppler shift the user is experiencing, and Doppler shift disturbs flatness of the subcarrier in frequency. To keep the narrow-band subcarrier bandwidth flat, the subcarrier width is fixed. For higher bandwidth, the number of carriers are increased but subcarrier spacing is constant.

This introduces a flexible architecture and makes OFDM agnostic to phase noise and Doppler shifts, since narrower subcarrier spacing decreases the performance, and transcending this issue requires expensive and complex designs.

For example, OFDMA mode of IEEE 802.16e-2005, which is expected to operate between 2–6 GHz, considers mobility at a maximum speed of 125 km/h (35 m/s). Doppler shift for operation in 2.5 GHz carrier frequency (f_c) is

$$f_\mathrm{m} = \frac{v f_\mathrm{c}}{c} = \frac{35\,\mathrm{m}/s\,2.510^9\,\mathrm{Hz}}{3x10^8\,\mathrm{m/s}} = 291\,\mathrm{Hz} \tag{5.1}$$

and it is 408 Hz when $f_\mathrm{c} = 3.5\,\mathrm{GHz}$ and 700 Hz when $f_\mathrm{c} = 6$ GHz. When maximum Doppler shift is considered, coherence time of the channel is $T_\mathrm{C} = \sqrt{\frac{9}{16.\pi.f_\mathrm{m}^2}} = 1.1\,\mathrm{ms}$. This would require an update rate of $\approx$1KHz for channel estimation and equalization. The delay spread (T_S) for mobile environment is 20 μs specified by The International Telecommunications Union (ITU-R).[2] Associated coherence bandwidth (B_C) is

$$B_\mathrm{C} \approx \frac{1}{5.T_\mathrm{S}} = \frac{1}{5.20\,\mu\mathrm{s}} = 10\,\mathrm{KHz}, \tag{5.2}$$

where sought frequency correlation is 50%. This means that over a 10-KHz subcarrier width, the fading is considered flat.

SOFDMA subcarrier spacing is independent of channel bandwidth. Scalability ensures that system performance is consistent across different RF channel sizes (1.25–20 MHz).

Larger FFT sizes of SOFDMA can cope with larger delay spreads, making the technology more resistant to multipath fading that is characteristic of NLOS propagation, particularly with larger RF channels.

[1] The scalable OFDMA is introduced in the IEEE 802.16 WirelessMAN OFDMA mode together with other features such as AMC subchannels, hybrid automatic repeat request (HARQ), high-efficiency uplink (UL) subchannel structures, multiple input multiple output (MIMO) diversity, enhanced advanced antenna systems (AAS), and coverage enhancing safety channels to simultaneously enhance coverage and capacity of mobile systems with tools to tune between mobility and capacity.

[2] "The worst case delay spread is 5.24 μs specified by SUI-6 of Stanford University Interim (SUI) channel model for Terrain Type A which is a hilly terrain with moderate-to-heavy tree densities..."

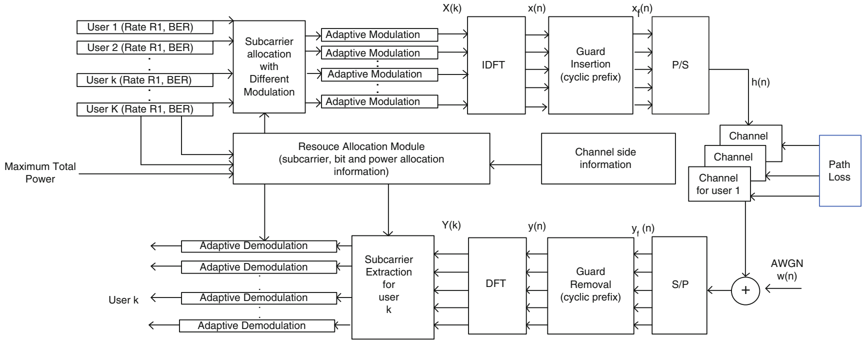


Fig. 5.10 Orthogonal frequency division multiple access system

5.3.2 System Model

This section outlines the OFDMA system seen in Fig. 5.10 and states the resource allocation problem. Unlike in a point-to-point OFDM system, K users are involved in the OFDMA system to share N subcarriers. The difference arises in the forming and deforming of FFT block. The rest is the same as in an OFDM system. Each user allocates nonoverlapping set of subcarrriers S_k, where the number of subcarriers per user is $J(k)$. We use $X_k(l)$ to denote the lth subcarrier of the FFT block belonging to kth user. $X_k(l)$ is obtained by coding the assigned bits c with the corresponding modulation scheme. In the downlink, the $X_k(l)$ are multiplexed to form the OFDM symbol[3] of length $(N+L)$ with the appended guard prefix L in order to eliminate ISI. At the uplink, the OFDM symbol is formed in the base station as follows:

$$x(l) = \sum_{k=0}^{K-1} \sum_{n=0}^{J(k)-1} X_k(n) e^{j\frac{2\pi}{N}(I_k(n))l}, \tag{5.3}$$

with $n = -L, \ldots, N-1$, where $I_k(n)$ denotes the subcarrier assigned to the kth user. A resource allocation problem comes into the picture when associating the set of subcarriers to the users with different bits loaded into them. The received signal from the jth user is:

$$y_j(l) = x(l) \bigotimes h_j(l) + w(l), \tag{5.4}$$

where $h_j(t)$ is the baseband impulse response of the channel between base station (BS) and jth user. Equation (5.4) is the received signal $y(t)$ sampled at rate $1/T$. The first L samples are discarded and N-point FFT is computed. Consequently, the data of the jth user is

$$Y_j(n) = \begin{cases} X_j(n)H_j(i_j(n)) + W(n), & \text{if } i_j(n) \in S_j \\ 0, & \text{otherwise} \end{cases}, \tag{5.5}$$

where $H_j(n) := \sum_i h_j(i) \exp(j2\frac{\pi}{N}ni)$ is the frequency response of the channel of kth user.

[3] An OFDMA symbol is defined as one OFDM FFT block.

5.3.3 QoS Awareness

In a perfectly synchronized system, the allocation module of the transmitter assigns subcarriers to each user according to some QoS criteria.

Each user when creating a service flow specifies the type of the service flow. Typically a service type is characterized by data rate and bit error rate (BER). When the service flow is created, it basically has provisioned a set of service flow parameters that are *requested* from the scheduler in MAC layer. These informations are retrieved by base station from the network and sent to the scheduler in the MAC layer.

The MAC layer checks if such a *requested*-QoS$_k$ is feasible or not for user k, and determines the *offered*-QoS$_k$ according to its scheduling algorithm.

The type of flows may differ according to their flow parameters. They may coarsely fall into two categories: real-time resource allocation and non–real-time resource allocation. Real-time resource allocation has fixed data rate and BER requirements, in which the scheduler tries to satisfy these for each user subject to given fixed power constraint. This type of resource allocation differs from resource allocation from a capacity standpoint. “Water-filling” or “water-pouring” strategy derived by Gallager in 1968 allocates more power to better channels to achieve Shannon capacity. Consequently, if the channel is bad for a long time, the system will wait until it gets better. As a result, a user might not be scheduled for its bad channel for a long time.

Non–real-time resource allocation suits for best effort traffic, where the flow has burst nature. In this case, proportional fairness is preferred to satisfy certain type of fairness as well as proportional resource allocation according to requested demands.

5.3.4 Channel

The power level of the modulation is adjusted to overcome the fading of the channel. The transmission power for AWGN channel can be predicted. In addition, the channel gain of subcarrier n to the corresponding user k should be known. The channel gain of the subcarrier is defined as:

$$\alpha_{k,n} = H_k(n) * \mathrm{PL}_k, \tag{5.6}$$

where PL is the pathloss, defined by:

$$\mathrm{PL}_k = \mathrm{PL}(d_o) + 10\alpha \log_{10}(d_k/d_o) + \chi_\sigma, \tag{5.7}$$

where d_o is the reference distance, d_k is the distance between the transmitter and the receiver, α is the pathloss component and χ_σ is a Gaussian random variable for shadowing with a standard deviation σ. An example of channel gain can be seen in Fig. 5.11.

The channel information is assumed to be known at the transmitter and receiver. The channel may be assumed to be reciprocal: BS is able to estimate the channel

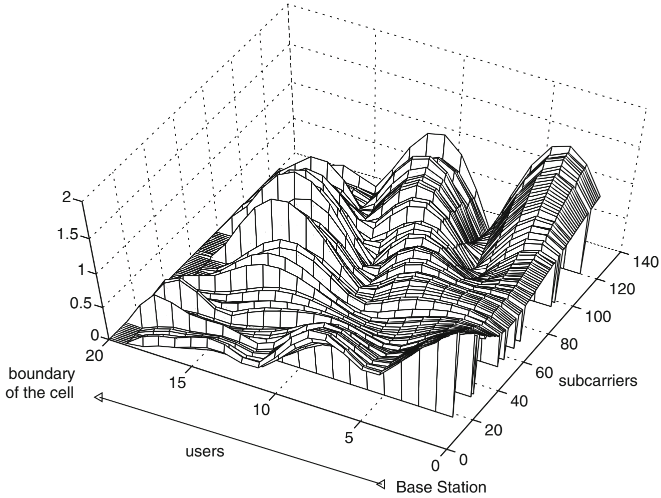


Fig. 5.11 An example of channel gain

of all BS-to-mobile links based on the received uplink transmission as long as the channel variation is slow. As a result, the resource allocation should be done within the coherence time.

5.4 Subcarrier Allocation: Fixed QoS Constraints

With the channel information, the objective of resource allocation problem can be defined as maximizing the throughput subject to a given total power constraint regarding the user's QoS requirements.

As we clarify further, BER_k of the transmission should not be higher than the required BER_k, and the data rate of every user should be equal to the requirement R_k.

Let us define $\gamma_{k,n}$ as the indicator of allocating the nth subcarrier to the kth user. The transmission power allocated to the nth subcarrier of kth user is expressed as:

$$P_{k,n} = \frac{f_k(c_{k,n}, \mathrm{BER}_k)}{\alpha_{k,n}^2}, \tag{5.8}$$

where $f_k(c_k, n)$ is the required received power with unity channel gain for reliable reception of $c_{k,n}$ bits per symbol.

We can formulate the resource allocation problem with an imposed power constraint as:

$$\begin{aligned} &\max\nolimits_{c_{k,n},\gamma_{k,n}} R_k = \textstyle\sum_{n=1}^{N} c_{k,n}\gamma_{k,n} \quad \text{for all k} \\ &\text{subject to } P_T = \textstyle\sum_{k=1}^{K}\sum_{n=1}^{N} \frac{f_k(c_{k,n},\mathrm{BER}_k)}{\alpha_{k,n}^2}\gamma_{k,n} \le P_{\max}, \end{aligned} \tag{5.9}$$

where the limit on the total transmission power is expressed as P_{Max} for all $n \in \{1,...,N\}$, $k \in \{1,...,K\}$, and $c_{k,n} \in \{1,...,M\}$.

If there is no power constraint, (5.9) is changed in order to minimize P_T subject to allocating R_k bits for all k (i.e., problem is to find the values of the $\gamma_{k,n}$ and the corresponding $c_{k,n}$ while minimizing P_T). As it can be seen, the cost function is the power consumption matrix in (5.8). Rather than using $\alpha^2_{k,n}$, we adopt $P_{k,n}$, since in this case, modulation type and BER are involved in the decision process.

This formulation can be modified for uplink resource allocation, in which each user has a constraint for the individual transmit power.

5.4.1 Optimal Solution

In a multiuser environment with multiple modulation techniques, the solution to the problem is complicated since the optimal solution needs to pick the subcarriers in balance. We can classify the problem according to each set of bits assigned to a subcarrier.

For user k, $f_k(c_{k,n})$ is in $\{f_k(1,\text{BER}_k),\ldots,f_k(M,\text{BER}_k)\}$. We can construct M times $[K \times N]$ power matrices$\{\text{P}^{\text{c}}\}$ for each c. For a constant c, $\{f(c)\}$ can be computed, and the transmission power requirement can be found with (5.8). The dimension of the indicator function is incremented and represented by $\gamma_{k,n,c}$ as follows:

$$\gamma_{k,n,c} = \begin{cases} 1, & \text{if } c_{k,n} = c \\ 0, & \text{otherwise} \end{cases}. \tag{5.10}$$

The above problem can be solved with integer programming (IP). We refer to the IP approach as the optimal solution to the resource allocation problem. The nonlinear approximation requires more computation than the IP.

There are $K \times N \times M$ indicator variables and M power matrices where the entries of each matrix for a given c can be found from:

$$P^c_{k,n} = \frac{f_k(c,\text{BER}_k)}{\alpha^2_{k,n}}. \tag{5.11}$$

Using (5.11) as an input, the cost function can be written as

$$P_T = \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{c=1}^{M} P^c_{k,n} \gamma_{k,n,c} \tag{5.12}$$

and the description of the IP problem is:

$$\min_{\gamma_{k,n,c}} P_T, \quad \text{for } \gamma_{k,n,c} \in \{0,1\} \tag{5.13}$$

subject to:

$$R_k = \sum_{n=1}^{N} \sum_{c=1}^{M} c.\gamma_{k,n,c}, \quad \text{for all k,}$$

and

$$0 \leq \sum_{k=1}^{K} \sum_{c=1}^{M} \gamma_{k,n,c} \leq 1, \quad \text{for all } n.$$

Although the optimal solution gives the exact results, from an implementation point of view, it is not preferred, since in a time varying channel, in order to allocate the subcarriers within the coherence time, the allocation algorithm should be fast and the IP complexity increases exponentially with the number of constraints. This real-time requirement leads to searching suboptimal solutions that are fast and close to the optimal solution. Several suboptimal allocation schemes are proposed for different settings in the literature. Up to now suboptimal solutions differ in the modulation type. There are a few suboptimal schemes that use adaptive modulation; the rest assume fixed modulation.

5.4.2 *Suboptimal Solutions*

In most attempts to simplify the resource allocation problem, the problem is decomposed into two procedures: a subcarrier allocation with fixed modulation, and bit loading. Subcarrier allocation with fixed modulation deals with one P^c matrix with fixed c, and then by using bit loading scheme, the number of bits is incremented.

5.4.3 *Subcarrier Allocation*

We know that $f_k(x,y)$ is a convex function. We can start with $P^1_{k,n}$ and we can define new $\bar{R}_k$ with $\sum_{k=1}^{K} \bar{R}_k \leq N$, which can be obtained by decrementing R_k properly. Then the solution to this problem can be solved with linear programming (LP) or mapping to the Hungarian problem. Although the Hungarian algorithm is proposed as an optimal solution for resource allocation with a fixed modulation, we consider it as a suboptimal solution for adaptive modulation.

5.4.3.1 Linear Programming

Resource allocation problem formulated by using LP is as follows:

$$P_T = \min \sum_{k=1}^{K} \sum_{n=1}^{N} P^1_{k,n} \rho_{k,n}, \tag{5.14}$$

where $\rho_{k,n} \in [0,1]$ and the constraints become:

$$\begin{array}{ll} \sum_{n=1}^{N} \rho_{k,n} = \bar{R}_k & \forall\, k \in \{1,...,K\} \\ \sum_{k=1}^{K} \rho_{k,n} = 1 & \forall\, n \in \{1,...,N\}. \end{array}$$

After LP, the $[K \times N]$ allocation matrix has entries ranging between 0 and 1. The entries are converted to integers by selecting the highest $\bar{R}_k$ nonzero values from N columns for each k and assigning them to the kth user.

5.4.3.2 Hungarian Algorithm

The problem described above can also be solved by an assignment method such as Hungarian algorithm. The Hungarian algorithm works with square matrices. Entries of the square matrix can be formed by adding $\bar{R}_k$ times the row of each k. The problem formulation is:

$$P_T = \min \sum_{k=1}^{N} \sum_{n=1}^{N} P_{k,n}^{1} \rho_{k,n}, \tag{5.15}$$

where $\rho_{k,n} \in \{0,1\}$ and the constraints become:

$$\begin{array}{ll} \sum_{n=1}^{N} \rho_{k,n} = 1 & \forall\, n \in \{1,\dots,N\} \\ \sum_{k=1}^{N} \rho_{k,n} = 1 & \forall\, k \in \{1,\dots,N\}. \end{array}$$

Although the Hungarian method has computation complexity $O(N^4)$ in the allocation problem with fixed modulation, it may serve as a base for adaptive modulation.

5.4.4 *Bit Loading Algorithm*

The bit loading algorithm (BLA) is performed after the subcarriers are assigned to users that have at least $\bar{R}_k$ bits assigned. Bit loading procedure is as simple as incrementing bits of the assigned subcarriers of the users until $P_T \leq P_{\max}$. Define $\Delta P_{k,n}(c)$ as the additional power needed to increment one bit of the nth subcarrier of kth user,

$$\Delta P_{k,n}(c_{k,n}) = \frac{[f(c_{k,n}+1) - f(c_{k,n})]}{\alpha_{k,n}^2}. \tag{5.16}$$

The bit loading algorithm assigns one bit at a time with a greedy approach to the subcarrier. Representation is as follows: $\{\arg\min_{k,n} \Delta P_{k,n}(c_{k,n})\}$.

BL Algorithm

```
STEP 1: For all $n$, Set $c_{k,n} = 0$, $\Delta P_{k,n}(c_{k,n})$, and $P_T = 0$;
STEP 2: Select $\bar{n} = \arg\min_n \Delta P_{k,n}(0)$;
STEP 3: Set $c_{k,\bar{n}} = c_{k,\bar{n}} + 1$ and $P_T = P_T + \Delta P_{k,n}(c_{k,n})$;
STEP 4: Set $\Delta P_{k,n}(c_{k,\bar{n}})$;
STEP 5: Check $P_T \leq P_{\max}$ and $R_k$ for $\forall k$, if not satisfied GOTO STEP 2.
STEP 6: END.
```

It is a simple algorithm. Bits on the subcarriers are incremented one by one. If there is no power constraint, procedure runs for $\sum_{k=1}^{K} R_k$ times. This algorithm enables us to convert the fixed modulation schemes into adaptive modulation.

The Hungarian approach and LP approach with bit loading appear as two different suboptimal solutions to the resource allocation with adaptive modulation. Let us call them GreedyHungarian and GreedyLP, respectively and look at more iterative and fair schemes.

5.4.5 Iterative Solution

The GreedyLP and GreedyHungarian methods both first determine the subcarriers and then increment the number of bits on them according to the rate requirements of users. This may not be a good schedule in certain cases. For instance, consider a user with only one good subcarrier and a low rate requirement. The best solution for that user is allocating its good carrier with high number of bits. But if GreedyLP or GreedyHungarian is used, the user may have allocated more than one subcarrier with lower number of bits, and in some cases, its good subcarrier is never selected. Consider another scenario where a user does not have any good subcarrier (i.e., it may have a bad channel or be at the edge of the cell). In this case, rather than pushing more bits and allocating less subcarriers as in GreedyLP and GreedyHungarian, the opposite strategy is preferred since fewer bits in higher number of subcarriers give better result. Another difficulty arises in providing fairness. Since GreedyLP and GreedyHungarian are based on greedy approach, the user in the worst condition usually suffers. In any event, these are complex schemes and simpler schemes are needed to finish the allocation within the coherence time.

Iterative solution introduces a simple, efficient, and fair subcarrier allocation scheme. Iterative solution is composed of two modules named *scheduling* and *improvement* modules. In the scheduling step, bits and subcarriers are distributed to the users and passed to the improvement module where the allocation is improved iteratively by bit swapping and subcarrier swapping algorithms.

5.4.6 Fair Scheduling Algorithm

Fair scheduling algorithm (FSA) is a simple and mixed allocation scheme that considers fair allocation among users with adaptive modulation. The allocation procedure starts with the highest level of modulation scheme. In this way, it tries to find the best subcarrier of a user to allocate the highest number of bits.[4] With this

[4] We can describe the strategy by an analogy: “The best strategy to fill a case with stone, pebble and sand is as follows. First filling the case with the stones and then filling the gap left from the stones with pebbles and in the same way, filling the gap left from pebbles with sand. Since filling in opposite direction may leave the stones or pebbles outside.”

strategy, more bits can be allocated and the scheme becomes immune to uneven QoS requirements. The FSA runs a greedy release algorithm (GRA) if there are non allocated subcarriers after the lowest modulation turn and the rate requirement is not satisfied. GRA decrements one bit of a subcarrier to gain power reduction, which is used to assign higher number of bits to the users on the whole. FSA is described as follows:

FS Algorithm

```
STEP 1: Set $c = M$, Select a $k$, and $P_T = 0$;
STEP 2: Find $\bar{n} = \arg\min_n P^c_{k,n}$;
STEP 3: Set $R_k = R_k - c$ and $\rho_{k,\bar{n}} = 1$, Update $P_T$, Shift to the next $k$;
STEP 4: If $P_T > P_{\max}$, Step Out and Set $c = c - 1$, GOTO STEP 2.
STEP 5: If $\forall k$ ,$R_k < c$, Set $c = c - 1$, GOTO STEP 2.
STEP 6: If $\{c == 1\}$,$\sum_{k=1}^{K}\sum_{n=1}^{N}\rho_{k,n} < N$, $P_T > P_{\max}$, Run "Greedy Release"
        and GOTO STEP 2.
STEP 7: END.
```

5.4.7 Greedy Release Algorithm

Greedy release algorithm (GRA) tends to fill the unallocated subcarriers. It releases one of the bits of the most expensive subcarrier to gain power reduction in order to drive the process. GRA works in the opposite direction of BLA. GRA is described as follows:

GR Algorithm

```
STEP 1: Find $\{\bar{k}, \bar{n}, \bar{c}_{\bar{k},\bar{n}}\} = \arg\max_{k,n,c} P^c_{k,n}\rho_{k,n}\ \forall c$;
STEP 2: Set $\bar{c}_{k,n} = \bar{c}_{k,n} - 1$, $P_T = P_T - \Delta P_{\bar{k},\bar{n}}(c_{\bar{k},\bar{n}})$;
STEP 3: Set $c = c_{\bar{k},\bar{n}} - 1$;
STEP 4: Finish.
```

5.4.8 Horizontal Swapping Algorithm

The horizontal swapping algorithm (HSA) aims to smooth the bit distribution of a user. When the subcarriers are distributed, the bit weight per subcarrier can be adjusted to reduce power. One bit of a subcarrier may be shifted to the other subcarrier of the same user if there is a power reduction gain. Therefore, variation of the power allocation per subcarrier is reduced and a smoother transmission is performed. HSA is described as follows:

HS Algorithm

```
STEP 1:  Set $P_C = \infty$
STEP 1a: Find $\{\bar{k}, \bar{n}, \bar{c}_{\bar{k},\bar{n}}\} = \arg\max_{k,n,c}(P^c_{k,n}\rho_{k,n}) < P_C\ \forall c$;
STEP 2:  Define $n \in S_k$, where $\{\rho_{k,n} == 1\}$ for $\forall n$;
STEP 3:  Set $\Delta_{\dot{n}} = \max_n \Delta P_{\bar{k},\bar{n}}(c_{\bar{k},\bar{n}} - 1) - \Delta P_{\bar{k},\dot{n}}(c_{\bar{k},\dot{n}})$, $\dot{n} \in S_k$;
STEP 4:  Set $P_C = P^{\bar{c}}_{\bar{k},\bar{n}}$;
STEP 4a: if $\Delta_{\dot{n}} > 0$, Set $P_T$=$P_T - \Delta_{\dot{n}}$
STEP 4b: Set $c_{\bar{k},\bar{n}} = c_{\bar{k},\bar{n}} - 1$, $c_{\bar{k},\dot{n}} = c_{\bar{k},\dot{n}} + 1$ GOTO Step 1a;
STEP 5:  if $\{P_C == \min_{k,n,c}(P^c_{k,n}\rho_{k,n})\}$, END.
```

5.4.9 Vertical Swapping Algorithm

Vertical swapping is done for every pair of users. In each iteration, users try to swap their subcarriers such that the power allocation is reduced. There are different types of vertical swapping. For instance, in triple swapping, user *i* gives its subcarrier to user *j* and in the same way user *j* to user *k* and user *k* to user *i*.

Pairwise swapping for fixed modulation can be based on power or channel gain as a decision metric. For adaptive modulation, there is more than one class where each class is defined with its modulation (i.e., number of bits loaded to a subcarrier) and swapping is only within the class. Each pair of users swap their subcarriers that belong to the same class if there is a power reduction. In this way, adjustment of subcarrier is done across users in order to converge to the optimal solution. VSA is described as follows:

VS Algorithm

```
STEP 1:  $\forall$ pair of user $\{i, j\}$;
STEP 1a: Find $\partial P_{i,j}(n) = P^{\dot{c}}_{i,n} - P^{\dot{c}}_{j,n}$ and $\Delta^{\hat{n}} P_{i,j} = \max \partial P_{i,j}(n)$, $\forall n \in S_i$;
STEP 1b: Find $\partial P_{j,i}(n) = P^{\dot{c}}_{j,n} - P^{\dot{c}}_{i,n}\ \Delta^{\breve{n}} P_{j,i} = \max \partial P_{j,i}(n)$, $\forall n \in S_j$;
STEP 1c: Set $\Omega^{\hat{n},\breve{n}} P_{i,j} = \Delta^{\hat{n}} P_{i,j} + \Delta^{\breve{n}} P_{j,i}$;
STEP 1d: Add $\Omega^{\hat{n},\breve{n}} P_{i,j}$ to the $\{\Lambda\}$ list;
STEP 2:  Select $\Omega = \max_{(i,\hat{n}),(j,\breve{n})} \Lambda$;
STEP 3:  if $\Omega > 0$, Switch subcarriers and $P_T = P_T - \Omega$ GOTO STEP 1a;
STEP 4:  if $\Omega \leq 0$, END.
```

5.4.10 Performance Analysis

The performance of iterative algorithm has been compared with the proposed suboptimal GreedyHungarian and GreedyLP schemes and optimal IP scheme.

The M-ary quadrature amplitude modulation of 4-QAM, 16QAM, and 64QAM are adopted to carry two, four, or six bits/subcarrier. Required transmission power for c bits/subcarrier at a given BER with unity channel gain is:

$$f(c,\mathrm{BER}) = \frac{N_o}{3}\left[Q^{-1}\left(\frac{\mathrm{BER}}{4}\right)\right]^2 (2^c - 1), \tag{5.17}$$

where $Q^{-1}(x)$ is the inverse function of:

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-t^2/2} \mathrm{dt}.$$

The power spectral density level N_o is equal to unity, and gain of Rayleigh channel $E|H_k(n)|^2$ is also equal to unity. Number of subcarriers are 128 with a total transmission rate between 480 and 768 bits/symbol. BER requirement of users is selected from the list $\wp = \{1e-2, 1e-4\}$. In the simulations, depending on the constraint, either the rate requirements are fulfilled when the transmit power is minimized or the power constraint is fulfilled when rates are maximized. BER requirement, on the other hand, is satisfied in both situations. We distinguish the settings by naming them with or without power constraint.

Figure 5.12 shows the convergence of iterative scheme. In each iterative step, the power is reduced, keeping the total number of bits constant. The steepest decrease is observed in the HSA step, since the power reduction in bit swapping is higher

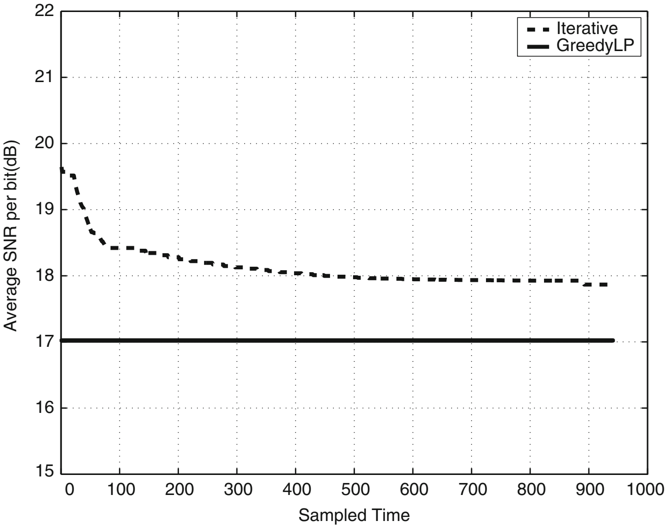


Fig. 5.12 Comparison of convergence of the iterative approach to the GreedyLP one

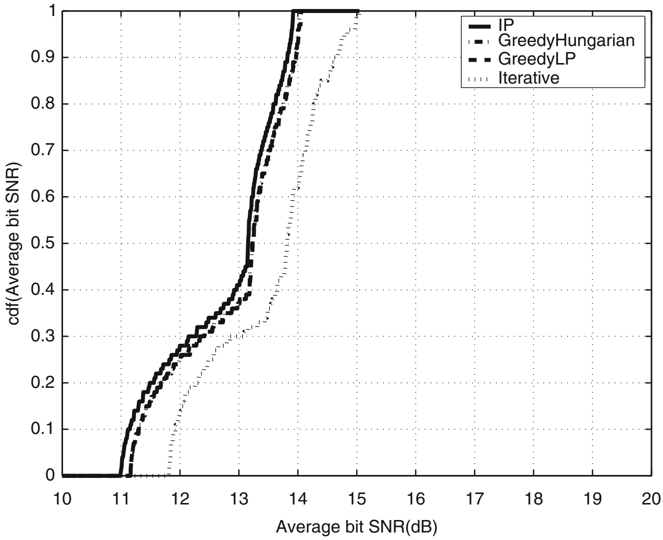


Fig. 5.13 Comparison of the cumulative distribution function of the average bit SNR (without power constraint)

than the one in subcarrier swapping because of the exponential growth of the $f(x,y)$ function. It can be seen from the figure that the iterative solution approximates the GreedyLP with time.

Figures 5.13 and 5.14 present the cumulative distribution functions of the average bit SNR for the cases without and with power constraints. There are four users in two sets of BER requirement, and each user has a rate requirement of 120 bits/symbol. It can be seen from the Fig. 5.13 that the iterative approach approximates the optimal solution up to 0.9 dB when there is no power constraint. When there is a power constraint, as seen in Fig. 5.14, the iterative approach outperforms the GreedyHungarian and GreedyLP approach and is close to the IP solution within 0.3 dB. The reason why iterative solution gives better performance than the suboptimal solution is its tight power control scheme, which allows transmission of higher number of bits. GRA is very important, since it decreases the variance of average bit SNR and makes the iterative approach perform better at the end by exchanging one high cost bit with more than one low cost bit, i.e., lower level modulation.

Figures 5.15 and 5.16 show the performance of the schemes in various channel fading and multiuser diversity situations. Figure 5.15 presents the average bit SNR as a function of root mean square (RMS) delay spread for different resource allocation schemes. As RMS delay spread increases, the fading variation increases, and so higher gains are obtained by adaptive allocation. We find that iterative approach is never more than 0.9 dB above the IP approach. Figure 5.16 shows the average bit SNR vs. the number of users where each has the same BER requirement of $1e-4$ and RMS value of 30 s. As the number of users increases, the probability of obtaining a good channel in the subcarriers increases. The iterative approach follows the lower bound within 0.9 dB and follows the GreedyHungarian and GreedyLP schemes within 0.5 dB.

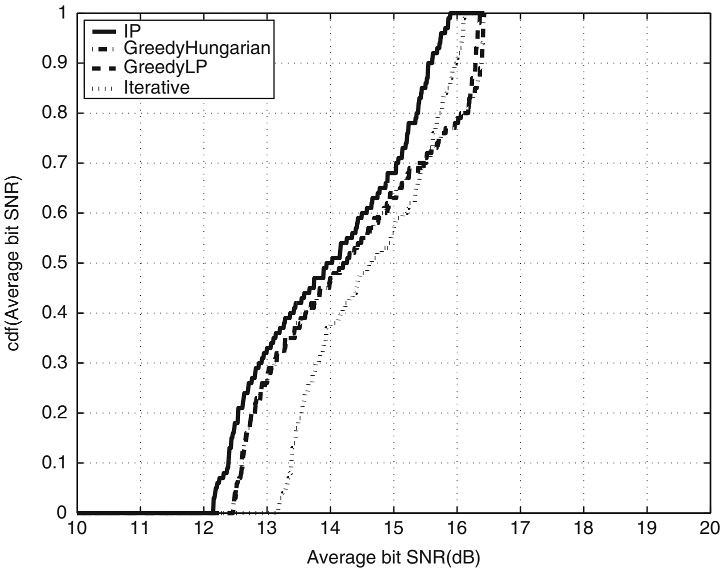


Fig. 5.14 Comparison of the cumulative distribution function of the average bit SNR (with power constraint)

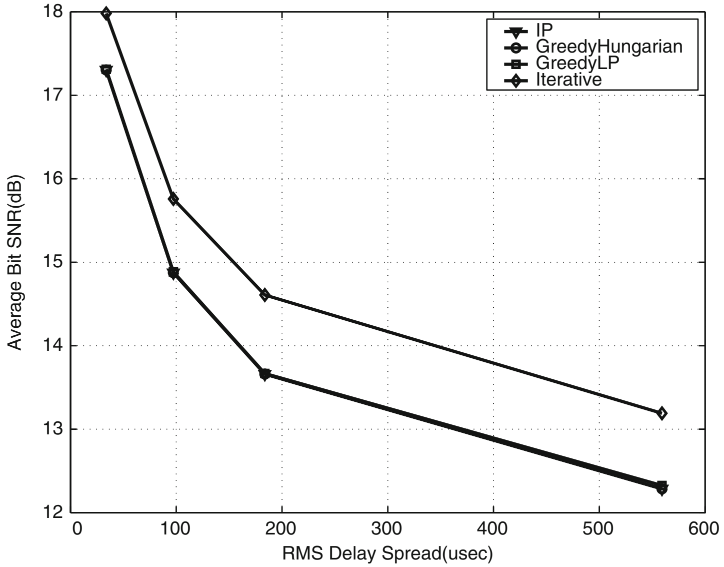


Fig. 5.15 Average bit SNR vs. channel fading and multiuser diversity: average vs. delay spread

Figures 5.17 and 5.18 show the standard deviation of the bit allocation of the users for different power constraints or different number of users. Each user has a BER requirement of $1e-4$ and the total transmit rate is 480 bits/symbol, which is equally distributed to each user. Each user has a 180 s RMS delay spread.

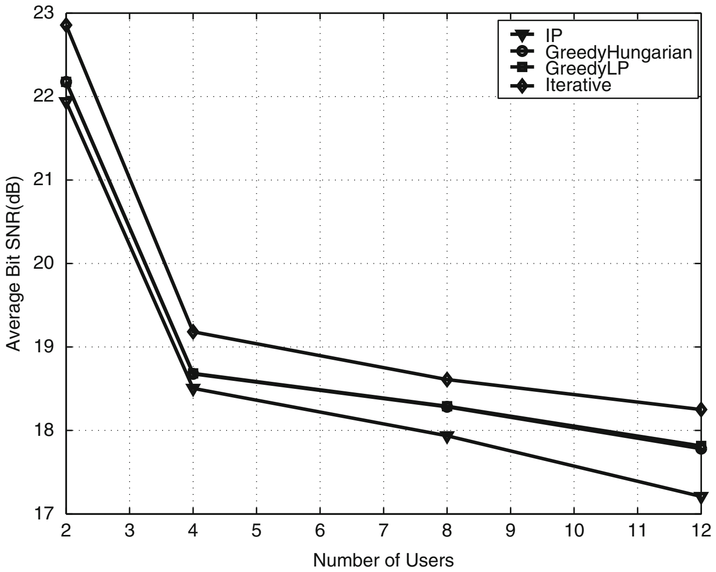


Fig. 5.16 Average bit SNR vs. channel fading and multiuser diversity: average bit SNR vs. number of users

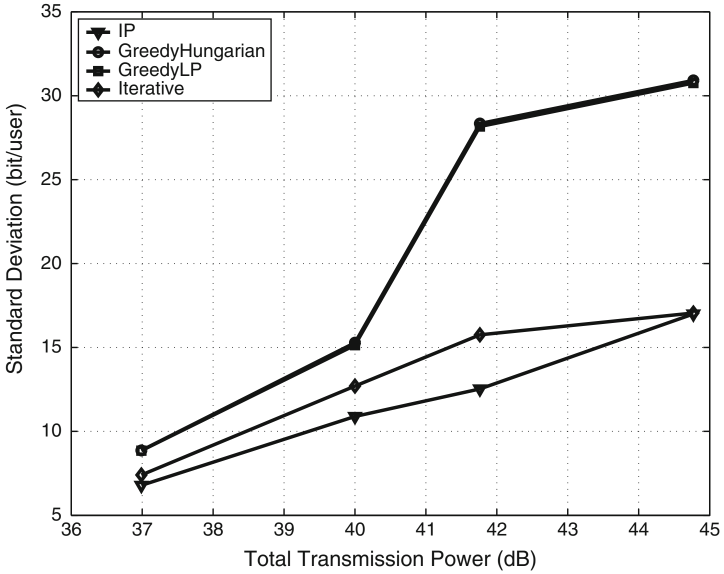


Fig. 5.17 Spectral efficiency vs. total power

Figure 5.17 presents the standard deviation of bit distribution among users under the total power constraint. It can be observed from the graph that iterative approach outperforms the GreedyHungarian and GreedyLP and is close to the IP. The FSA distributes the bits fairly compared with the greedy approach. The fairness property

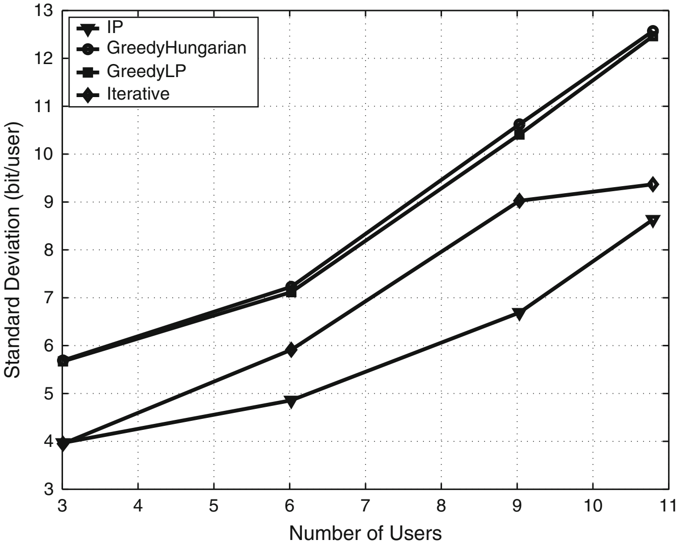


Fig. 5.18 Standard deviation of bits/user vs. number user

is an important metric for real-time data if there is tight power control. The iterative solution maintains fairness. As the total transmit power increases, the significance of a power control scheme decreases as can be inferred from the graph. In Figure 5.18 fairness is tested under varying number of users. The iterative approach again outperforms the GreedyHungarian and GreedyLP and closely follows the IP.

Figure 5.19 shows the average data rates per subcarrier vs. total power constraint when there are four users. Each user has a rate requirement of 192 bits/symbol (maximum rate) and BER requirement of $1e-4$. The performance of the iterative approach is close to that of the optimal and the difference between suboptimal and iterative approaches decreases as the total transmit power increases.

5.5 Subcarrier Allocation: Variable QoS

In the previous section, we considered the problem where the QoS requirements per symbol is fixed. Another way to approach resource allocation is in terms of capacity. Suppose there is no fixed requirements per symbol and the aim is to maximize capacity.

It has been shown that for point-to-point links, a fair allocation strategy maximizes total capacity and the throughput of each user in the long run, when the user's channel statistics are the same. This idea underlying the proposed fair scheduling algorithm exploits the multiuser diversity gain.

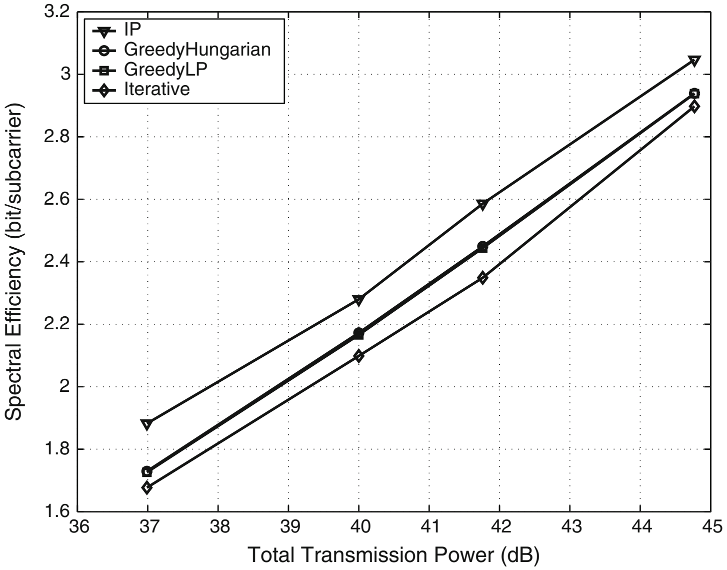


Fig. 5.19 Spectral efficiency vs. total transmission power

With a slight modification, we can extend the fair scheduling algorithm for point-to-point communication to an algorithm for point-to-multipoint communication. Suppose time-varying data rate requirement $R_k(t)$ is sent by the user to the base station as feedback of the channel condition. We treat symbol time as the time slot, and so t is discrete, representing the number of symbols. We keep track of average throughput $t_{k,n}$ of each user for a subcarrier in a past window of length t_c. The scheduling algorithm will schedule a subcarrier $\bar{n}$ to a user $\bar{k}$ according to the following criterion:

$$\{\bar{k},\bar{n}\} = \arg\max_{k,n} \frac{r_{k,n}}{t_{k,n}}, \tag{5.18}$$

where $t_{k,n}$ can be updated using an exponentially weighted low-pass filter. Here, we are confronted with determining the $r_{k,n}$ values. We can set $r_{k,n}$ to R_k/N, where N is the number of carriers. With this setting, the peaks of the channel for a given subcarrier can be tracked. The algorithm schedules a user to a subcarrier when the channel quality in that subcarrier is high, relative to its average condition in that subcarrier over the time scale t_c. When we consider all subcarriers, the fairness criterion match with the point-to-point case as follows:

$$\bar{k} = \max_{k} \frac{R_k}{T_k}, \tag{5.19}$$

where $T_k = \sum_{n=1}^{N} t_{k,n}$. It is the theoretical analysis of fairness property for point-to-point communication. We can apply these derivations for point-to-multipoint communication.

5.6 Frequency Reuse: Single Frequency Network

The resource allocation OFDMA for cellular network considers more than one cell as seen in Fig. 5.20. Each cell has a BS, and part of their coverage at the fringe is overlapped. Typically, a user at the fringe may have more than one choice. Flexibility of OFDMA can enable operation of cellular network with only one frequency.[5] This means that a set of subcarriers associated with the frequency can be reused across cells but not permitted to reuse within the same cell. And a user selects only one BS to be serviced from. This type of system removes the requirement for cell planning and called *single frequency network*.

As we discussed, scalable OFDMA maintains flat-fading and channel gain between BS_b and user_k to correspond frequency selective fading, pathloss, and shadowing is described by $\alpha_{k,n,b}$ for nth subcarrier. Additionally, there is co-channel interference $I_{k,n,b}$, since the subcarriers are reused. If B is the total number of base stations and K is the total number of users, $I_{k,n,b}$ is basically

$$I_{k,n,b} = \sum_{i \neq b}^{B} \sum_{j \neq k}^{K} \alpha_{k,n,i} P_{j,n,k}, \tag{5.20}$$

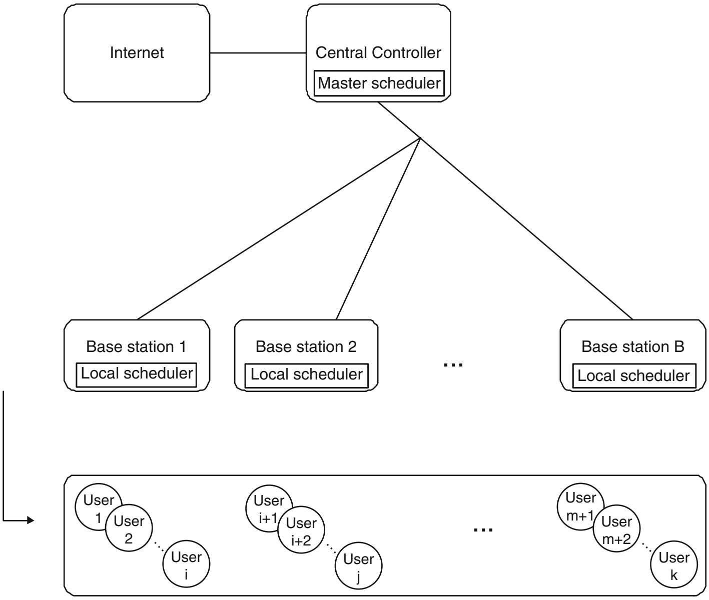


Fig. 5.20 Cellular OFDMA architecture

[5] Adapted from "OFDMA For Broadband Wireless Access," Pietrzyk, Artech House, Boston, 2006.

where $\alpha_{k,n,i}$ is the channel gain between interfering base station and the user. $P_{j,n,k}$ is the transmit power for the subcarrier n when transmitting to user j.

A subcarrier n is reused as long as $\text{SINR}_{k,n,b}$[6] satisfies the following equation;

$$\text{SINR}_{k,n,b} = \frac{\text{Received Power}}{\text{Interference} + \text{Noise}} = \frac{\alpha_{k,n,b} P_{k,n,b}}{I_{k,n,b} + \sigma^2} \geq \tau_{k,n,b}. \tag{5.21}$$

After rearrangement, we find

$$\alpha_{k,n,b} P_{k,n,b} - \tau_{k,n,b} \sum_{i \neq b, j \neq k} \alpha_{k,n,i} P_{j,n,i} \geq \sigma^2 \tau_{k,n,b}, \tag{5.22}$$

where this holds for every active cell b at subcarrier n assigned to user k. Equation (5.22) can be written in matrix form for each n as follows;

$$\mathbf{H}^n \mathbf{p}^n \geq \sigma^2 \tau^{\mathbf{n}}, \quad \mathbf{n} = \mathbf{1}, \cdots, \mathbf{N}, \tag{5.23}$$

where

$$\mathbf{H}^n = \begin{bmatrix} \alpha_{1,n,1} & -\tau_{1,n,2}\alpha_{1,n,2} & \cdots & -\tau_{1,n,\chi}\alpha_{1,n,\chi} \\ -\tau_{2,n,1}\alpha_{2,n,1} & \alpha_{2,n,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ -\tau_{\chi,n,1}\alpha_{\chi,n,1} & \cdots & \cdots & \alpha_{\chi,n,\chi} \end{bmatrix} \tag{5.24}$$

and

$$\mathbf{p}^n = [p_1^n, p_2^n, \cdots, p_\chi^n] \tag{5.25}$$

$$\tau^{\mathbf{n}} = [\tau_{\mathbf{1},\mathbf{n},\mathbf{1}}, \tau_{\mathbf{2},\mathbf{n},\mathbf{2}}, \cdots, \tau_{\chi,\mathbf{n},\chi}]^{\mathbf{T}}, \tag{5.26}$$

where χ is the number of base stations using subcarrier n and $\tau_{k,n,b}$ is equal to

$$\tau_{k,n,b} = \frac{N_o}{3}\left[Q^{-1}\left(\frac{\text{BER}_k}{4}\right)\right]^2 (2^c_{k,n,b} - 1). \tag{5.27}$$

Uncoded system is assumed with BER_k, which is bit error rate for user k and $c_{k,n,b}$ is the number of bits loaded for subcarrier n when it is allocated in base station b to user k. The feasible solution for $\mathbf{p}(n)$ is available if the solution is all positive for

$$\mathbf{H}^n \mathbf{p}^n = \sigma^2 \tau^{\mathbf{n}}, \quad \mathbf{n} = \mathbf{1}, \ldots, \mathbf{N} \tag{5.28}$$

and if b is using n and $k = \{k | \gamma_{k,n,b} = 1\}$ then $P_{k,n,b} = p_b^n$.

This is simply from the fact that if a user reduces the power, it causes less interference to others. An optimal solution for $\mathbf{p}(n)$ is when each user is at the minimal possible power so that SINR requirements is met with equality. Also convergence to optimal solution is possible with iterative schemes even the updates are asynchronous among users.

[6] SINR, Signal-to-interface-plus noise ratio.

5.6.1 Optimum Solution

Problem formulation for resource allocation in cellular OFDMA is similar to single-cell OFDMA. Optimization is considered with respect to either meeting the required quality with minimum power or maximizing the performance with given power. For fixed QoS, power is minimized:

$$\min \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{b=1}^{B} \gamma_{k,n,b} P_{k,n,b} \tag{5.29}$$

and for variable-QoS performance is maximized:

$$\max \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{b=1}^{B} \gamma_{k,n,b} c_{k,n,b}, \tag{5.30}$$

but they share the same constraints:

$$\begin{aligned} \mathrm{QoS}_u \le \mathrm{QoS}_o &= \textstyle\sum_{n=1}^{N} \sum_{b=1}^{B} \gamma_{k,n,b} c_{k,n,b}; \quad k = 1, \ldots, K, \\ P_{\max} \le oP_b^{\text{total}} &= \textstyle\sum_{k=1}^{K} \sum_{n=1}^{N} \gamma_{k,n,b} c_{k,n,b}; \quad b = 1, \ldots, B, \\ &\textstyle\sum_{k=1}^{K} \gamma_{k,n,b} \le 1; \quad n = 1, \ldots, N; \quad b = 1, \ldots, B, \end{aligned} \tag{5.31}$$

where $\gamma_{k,n,b} \in \{0, 1\}$. For only uplink power constraint is modified to

$$P_k^{\max} \ge oP_k^{\text{total}} = \sum_{n=1}^{N} \gamma_{k,n,b} P_{k,n,b}; \quad k = 1, \ldots, K; \quad b = 1, \ldots, B. \tag{5.32}$$

Optimum solution is a three-dimensional nonlinear and combinatorial allocation problem. There is a dependency on $P_{k,n,b}$, which is tightly coupled to allocations at co-channel cells. Optimum solution requires centrally coordinated implementation where the channel and traffic conditions are fed constantly.

In a cellular environment, there are independent cells serviced by base stations and a central controller who controls the signalling to base station for control and data communication. Central controller is a suitable place to run such a scheme but of course it is very challenging in a mobile environment, since channel and traffic conditions of the user may change over time, and reporting this information to central controller incurs delays.

An adaptive solution is much more suitable to deploy for mobile environments. Adaptive solution defines trigger mechanisms to execute the resource allocation and shares the decision between base station and central controller.

5.6.2 Adaptive Solution

Adaptive solution may be executed when one of the following conditions met

- New user
- Change in the channel condition

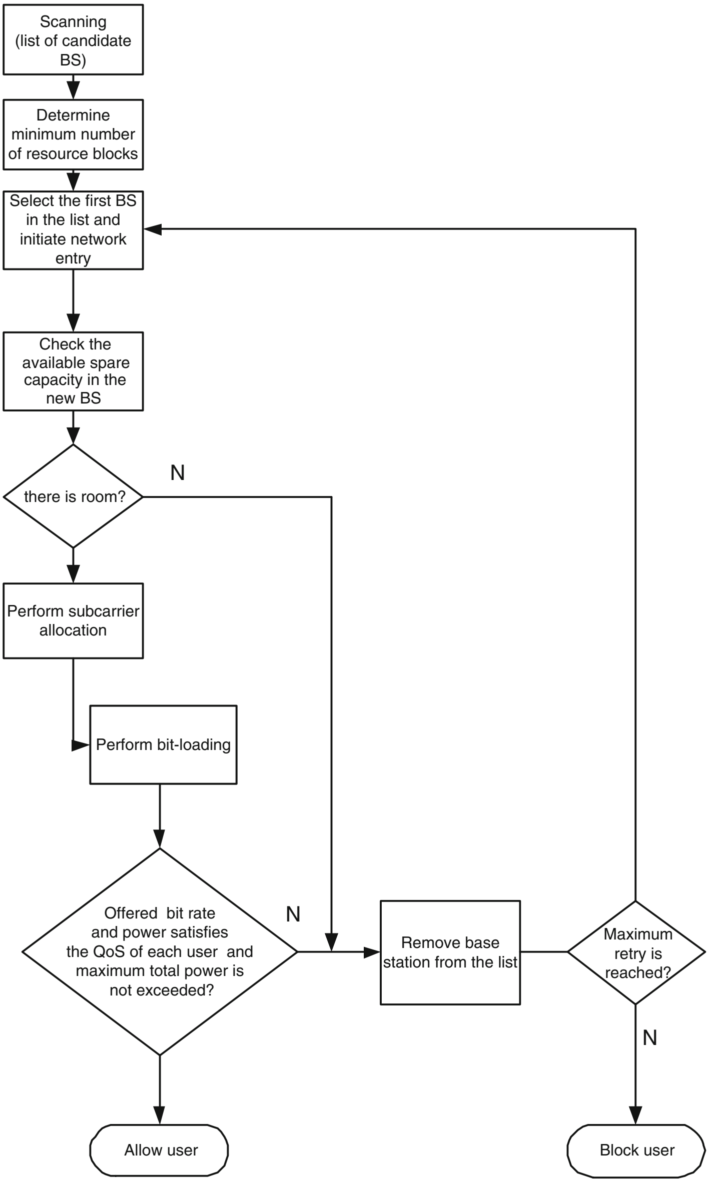

Scanning (list of candidate BS)
Determine minimum number of resource blocks
Select the first BS in the list and initiate network entry
Check the available spare capacity in the new BS
there is room?
N
Perform subcarrier allocation
Perform bit-loading
Offered bit rate and power satisfies the QoS of each user and maximum total power is not exceeded?
N
Remove base station from the list
Maximum retry is reached?
N
Allow user
Block user

Fig. 5.21 Adaptive solution

- Change in the traffic pattern
- User exit

A simple method is to consider only "new user" as a trigger mechanism and execute the adaptive solution in three steps as seen in Fig. 5.21:

1. Cell selection
2. Subcarrier allocation
3. Bit loading

5.6.2.1 Cell Selection

Cell selection in a typical cellular system is determined based on signal strength and available spare capacity of the cell for that particular user. A user identifies the available base stations and sorts them according to the preferred criteria. It picks the first base station in the list and initiates the network entry.

5.6.2.2 Resource Allocation

Resource allocation distributes the subcarriers within the cell after having found the best cell for the user. First, the minimum number of subcarriers required is distributed to the user and then the remaining subcarriers are allocated in distributed fashion to the users with least transmit power constraint. Minimum number of subcarriers are found for a user by dividing the required number of bits to number of bits carried by highest modulation, and the maximum number of subcarriers are found by dividing that to bits carried by lowest modulation. A user is restricted to occupy a number of subcarriers, which is less than equal to the maximum number of subcarriers needed.

A subcarrier is allocated to user if it leads to maximal power reduction among other users. After allocating the number of subcarriers, assignment of subcarriers to a particular index is performed by considering the channel gain matrix of user. Best channel gain is selected for that cell and subcarriers are assigned. Subcarriers can be swapped by horizontal and vertical swapping algorithm introduced for single- cell in the previous section.

5.6.2.3 Bit Loading

Bit-loading determines the power level and modulation for particular subcarrier along with co-channel cells. First, each subcarrier is assigned with minimum modulation. And then incremented to next higher modulation according to feasible power vector of (5.28). If there is no feasible vector than particular subcarrier for that user is not feasible in that cell. If there is no feasible subcarrier then the cell is saturated and there is no room for a new user.

Second, if there is a feasible solution then the subcarrier that incurs minimum power change is selected and assigned to the user. Of course it also changes the power allocation in co-channel cells as well. This loop repeats until data rate of all users belonging to the cell are satisfied without disrupting the users in the co-channel links.

Last, QoS criteria is checked: if met, the user is admitted to the system, otherwise it picks the next base station in the list and initiates network entry again. And excess subcarriers are released if bit loading does not use all the subcarriers.

Performance analysis shows that 0.6% subcarrier-reuse is achievable with 550 Mbps load and 19 cells over QPSK, 16QAM, and 64QAM. The total offered load saturates, since interference becomes blocking condition to prevent serving new users.

5.6.3 Heuristic Solution

Iterative solution also requires extensive communication among base stations and controller. Any time a new user joins network, power allocation in the co-channel cells also changes. The communication among base stations and central controller can be costly and comes with random latency, which disrupts the accuracy of scheduling.

An implementable way of utilizing the frequency reuse would move more decision to local scheduler in the BS and reduce the frequency of control information sent by the central controller.

Flexibility of an OFDMA frame in WiMAX allows to create segments in single frequency as in Fig. 5.22. In a typical cell deployment, three orthogonal frequency

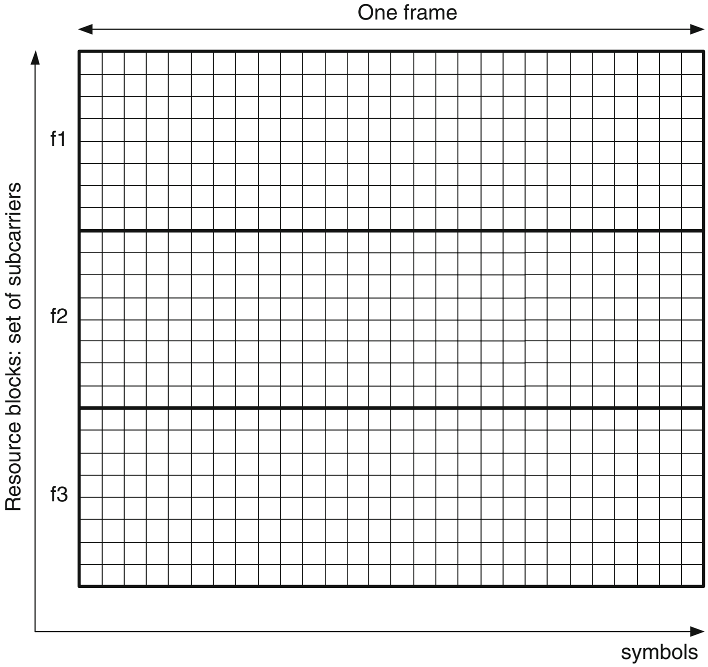


Fig. 5.22 Segmentation of an OFDMA frame

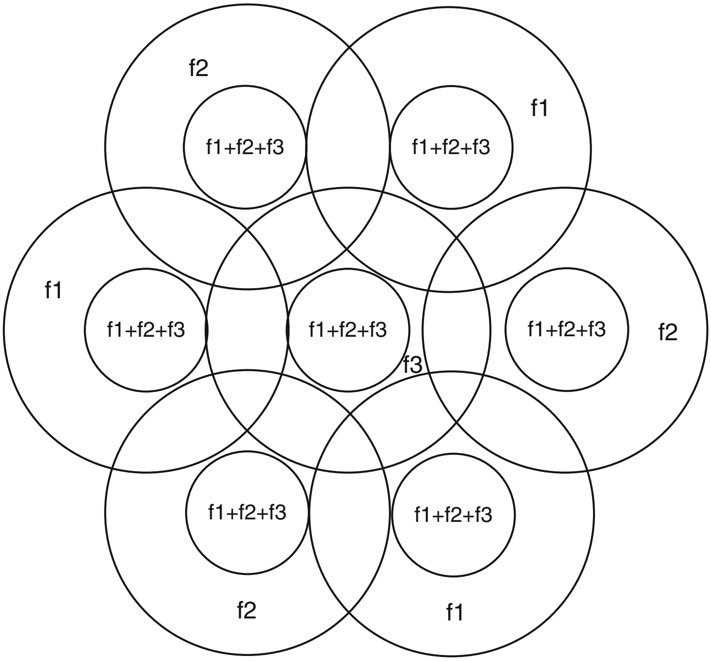


Fig. 5.23 Full reuse with reduced coverage in one-sector base station

$(f1, f2, f3)$ is adequate to achieve different frequency assignment in adjacent cells. Additionally, each cell can apply different transmit power level for each segment.

In one deployment, each cell can have one primary frequency and two secondary frequencies. Primary frequency is transmitted with full frequency and secondary frequencies are transmitted with just enough power to cover the nonoverlapping area of the cell. Figure 5.23 illustrates this deployment with omni-directional antennas and Fig. 5.24 depicts the case for three sector base stations.

Users (aka mobile stations, MS) are classified according to their viable links. A viable link has higher SNR then the minimum threshold for feasible communication between an MS and a BS. If an MS has only one viable link, then it is in the nonoverlapping region, otherwise it is in the overlapping region. MSs who are in the overlapping region are assigned to primary frequency and the MSs who are in the nonoverlapping region are assigned to one of the three frequencies. An MS scans periodically the channel to report its viable link status to BS.

This scheme performs well in downlink but in uplink co-channel interference is present, since the transmission of an MS reaches to other BSs when MS is in the overlapping area. Several schemes are present to reduce the interference in the uplink in such a configuration, but a more conservative deployment would eliminate this interference on the expense of reduced reuse.

In another deployment, within a frame flexible orthogonal channels are created in order to schedule MSs in the overlapping region. An illustration is shown in

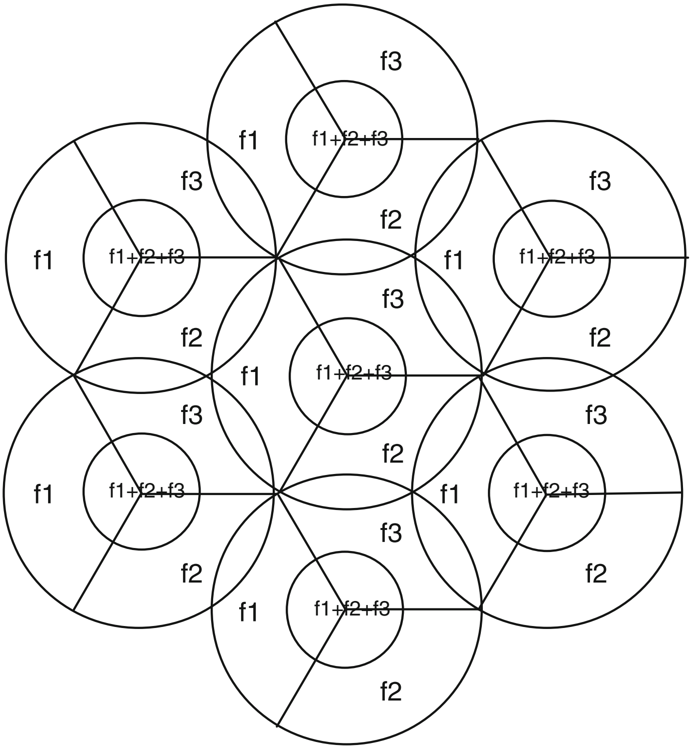


Fig. 5.24 Full reuse with power control in three-sector base station

Fig. 5.25. This scheme also operates with full power in all three segments but MSs are classified according to where they logically are. BSs operate in the following way: A BS is assigned a primary frequency and nonoverlapping (outer-cell) sections of the secondary frequencies. The BS blocks out the overlapping (inner-cell) sections of the secondary frequencies. If an MS is in the overlapping region, then it is scheduled in the primary frequency of the BS that is attached to, otherwise, an MS is scheduled to nonoverlapping section of one of three frequencies.

The ratio of nonoverlapping section to overlapping section is determined by the traffic pattern, and *index* value can be periodically updated with central controller. Within sections, local scheduler can utilize multiuser diversity in order to improve the efficiency. In this type of deployment, settings can be determined with fine tuning parameters in order to reduce the control signalling.

5.7 Code-Based Allocation: Flash-OFDM

More flexible single-frequency-network designs are also possible with minimal control signalling at the expense of less utilization. In flash-OFDM, considered in IEEE 802.20 standard, orthogonality of users within the cell is preserved by utilizing the

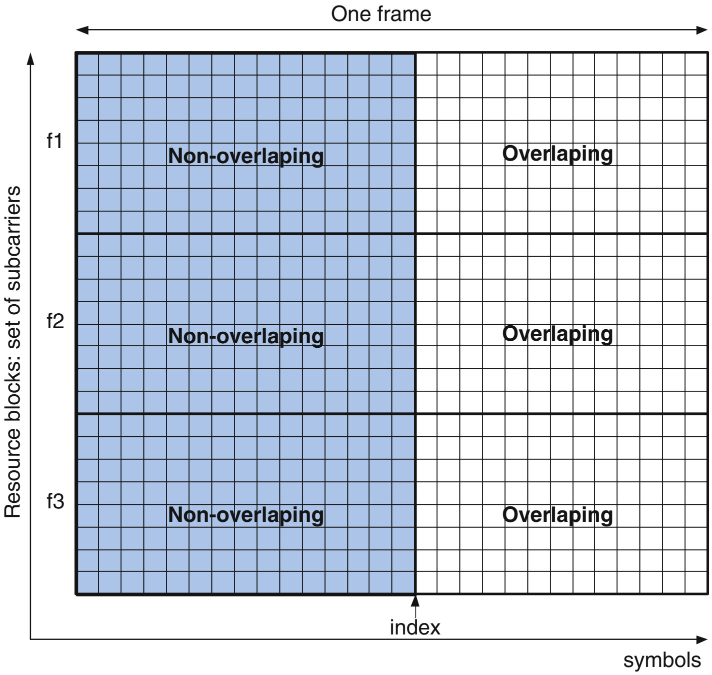


Fig. 5.25 Segmentation of an OFDMA frame for partial reuse without transmit power control

OFDM transmission and at the same time interference averaging is introduced as in CDMA systems. Interference averaging (aka interference diversity) addresses the intercell interference. Previously we saw that, with full reuse, two users in adjacent cells if they share the same subcarrier in an OFDM symbol interfere with each other. If they are close to each other, interference is severe. Introduced schemes above try to minimize these overlaps.

Flash-OFDM utilizes an hopping pattern in time and frequency in order to exploit interference diversity as in CDMA as well as frequency diversity. Interference diversity averages out the interference coming from more user, since interference coming solely from one user causes severe interference to each other. A hopping pattern is assigned to a user, which alternates the subcarriers at each symbol as seen in Fig. 5.26.

5.7.1 *Interference Diversity*

The link level outage depends on the aggregated interference coming from in-cell and out-of-cell. There is randomness associated with interference since interfering users have periods of activity, varying channel conditions, and imperfect power control. Due to this randomness, number of users in the system is soft-bounded for

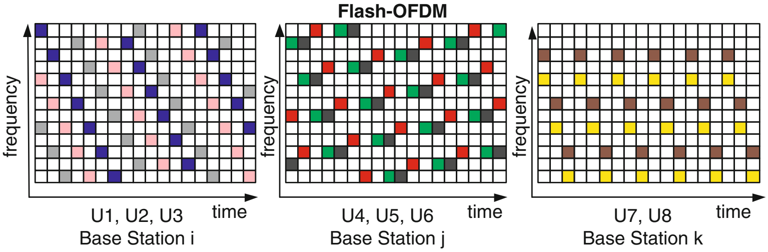


Fig. 5.26 Code-based hopping pattern for Flash-OFDM

Flash-OFDM as in CDMA based system. One notable difference is Flash-OFDM removes in-cell interference since allocated codes within the cell are orthogonal. The idea of interference diversity is similar to any diversity idea, interference diversity averages over the effects of different interferers.

Let us first look at the single cell in a CDMA system when there is perfect power control. Then, we simplify multi-cell system analysis for Flash-OFDM according to this analysis and cite how it differs against CDMA. If ε is chip power then SINR is

$$\mathrm{SINR}_{\mathrm{chip}} := \frac{\varepsilon_1}{\sum_{i\neq k} P_k + \sigma^2} \geq \tau, \quad k = 1,\ldots,K \tag{5.33}$$

and

$$\frac{GP_k}{\sum_{i\neq k} P_k + \sigma^2} \geq \tau, \quad k = 1,\ldots,K, \tag{5.34}$$

where $G = W/R$ is processing gain[7] and P_k is total received power of user k at the base station. If we rearrange and sum up all the inequalities, we get the following:

$$\left[\frac{G}{\tau} - \frac{K-1}{K}\right]\sum_{k=1}^{K} P_k \geq \sigma^2, \tag{5.35}$$

where we found $K < \frac{G}{\tau} + 1$ as a condition for feasible power vectors. If we substitute $G = W/R$, the spectral efficiency KR/W becomes

$$\frac{KR}{W} < \frac{1}{\tau} + \frac{1}{G}. \tag{5.36}$$

Since G is typically large, spectral efficiency[8] equals $1/\tau$.

If users are active with probability p then $\sum_{k=1}^{K} \upsilon_k < \frac{G}{\tau} + 1$, where υ_k is 1 when user k is active. Then maximum number of user in the network is bounded by $G/\tau + 1$ and outage probability becomes

[7] ‘Each information bit is modulated over G chips in CDMA and SINR per bit is G-fold of SINR per chip. Source : “Fundamentals of Wireless Communication by Tse and Viswanath.”

[8] For example, for IS-95 systems $\tau = 6\,$dB and maximum spectral efficiency is 0.25 bits/s/Hz.

$$\Pr\left[\sum_{k=1}^{K}\right] \leq p_{\text{out}}, \tag{5.37}$$

where the random variable $\sum \upsilon_k$ is binomially distributed with mean Kp and standard deviation $\sqrt{Kp(1-p)}$.

p_{out} when there are many users in the system can be approximated with Gaussian RV as follows;

$$p_{\text{out}} \approx Q\left[\frac{G/\tau + 1 - Kp}{\sqrt{Kp(1-p)}}\right] \tag{5.38}$$

and spectral efficiency ($\rho = KpR/W$) is

$$\rho = \frac{1}{\tau}\Delta, \tag{5.39}$$

where Δ is

$$\Delta = \left(1 + Q^{-1}\left(p_{\text{out}}\sqrt{\frac{1-p}{pK}} - \frac{1}{Kp}\right)\right)^{-1}. \tag{5.40}$$

As you can see Δ is a loss factor in spectral efficiency. This is due to the fact that there are fewer users in the system and if there are many users interference averaging would occur.

In a multicell setting, additionally we have out-of-cell interference in a CDMA system. This is also present in Flash-OFDM as well. We can also treat the subcarriers in Flash-OFDM as chips in CDMA system and look at the overall SINR:

$$\frac{GP_{k,i}}{\sum_{i\neq k}^{K_i} P_{k,i} + \sum_{b=1}^{B}\sum_{j}^{K_b} P_{j,i}}, \tag{5.41}$$

where $P_{j,i}$ is received power in cell i from user j residing in cell b. And K_i stands for the number of users in cell i.

If we look at very simple two cell structure as in Fig. 5.27,[9] which lays on one-dimensional canvas, separated with length d and each cell has uniformly distributed

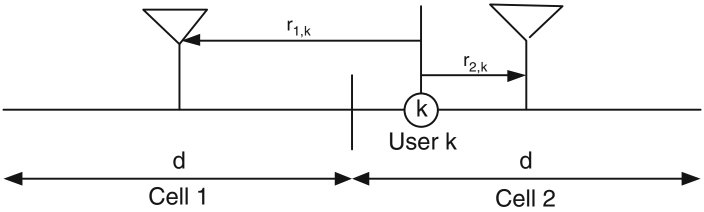


Fig. 5.27 Simplified two-cell structure on one-dimensional canvas

[9] Source: Exercise 4.11 in Fundamentals of Wireless Communication by Tse and Viswanath.

and perfectly power controlled K users. A user creates interference in the adjacent base station with $r^{-\alpha}$ power attenuation and there is no background noise in the system.

Following the assumptions, the SINR for a user connected to cell *one* is:

$$\frac{GP}{(K-1)P+\sum_k^K P(\frac{r_{2,k}}{r_{1,k}})^2} \geq \tau \tag{5.42}$$

then outage becomes

$$p_{\text{out}} = \Pr\left[\Lambda \geq \frac{G}{\tau} - (K-1)\right], \tag{5.43}$$

where Λ is $\sum_k^K (\frac{r_{2,k}}{r_{1,k}})^2$. Hence, the spectral efficiency $\rho = RK/\tau$ is

$$\rho = \frac{1}{\tau}\left[\frac{Q^{-1}(p_{\text{out}})}{2^{\sigma}\sqrt{K}}\left[\frac{1}{2\alpha+1} - \frac{1}{(1+\alpha)^2}\right]^{1/2} + \frac{1}{2^{\alpha}(\alpha+1)+1-\frac{1}{K}}\right]^{-1}, \tag{5.44}$$

where Λ can be approximated with a Gaussian RV with mean $\mu = K/(2^{\sigma}(\sigma+1))$ and standard deviation ς

$$\varsigma^2 = \frac{K}{2^{\sigma}}\left[\frac{1}{2\alpha+1} - \frac{1}{(\alpha+1)^2}\right]. \tag{5.45}$$

When K and W goes ∞, ρ plotted in Fig. 5.28 is

$$\lim_{K,W\to\infty}\rho = \frac{1}{\tau}\left[1 + \frac{1}{2^{\alpha}(\alpha+1)}\right]^{-1}. \tag{5.46}$$

Now, let us consider the Flash-OFDM where users are orthogonal in the cell. SINR for a user connected to cell one is

$$\frac{GP}{\sum_k^K P(\frac{r_{2,k}}{r_{1,k}})^2} \geq \tau \tag{5.47}$$

and spectral efficiency $\rho = RK/\tau$ is

$$\rho = \frac{1}{\tau}\left[\frac{Q^{-1}(p_{\text{out}})}{2^{\alpha}\sqrt{\frac{1}{2\alpha+1} - \frac{1}{(1+\alpha)^2}}} + \frac{1}{2^{\alpha}(\alpha+1)}\right]^{-1}. \tag{5.48}$$

As K and W go to ∞, we obtain

$$\lim_{K,W\to\infty}\rho = \frac{2^{\alpha}(\alpha+1)}{\tau} \tag{5.49}$$

as seen in Fig. 5.29.

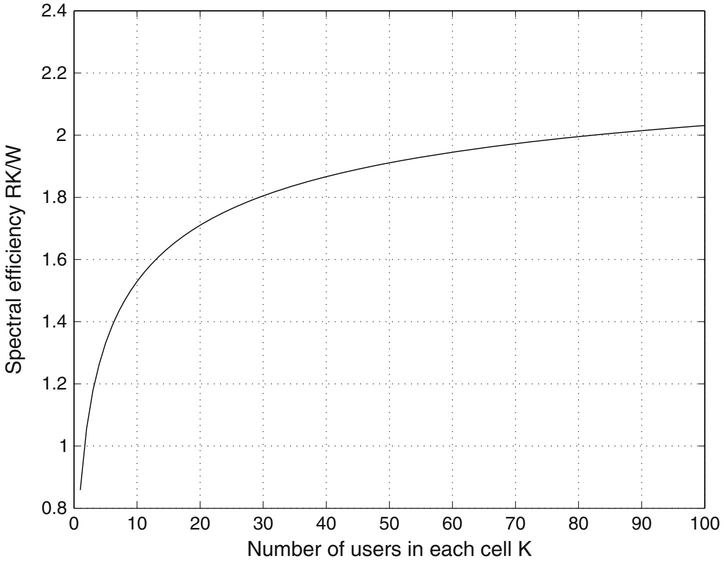


Fig. 5.28 Spectral efficiency with in-cell and out-of-cell interference: $Q^{-1}(p_{\text{out}}) = 2$, $\alpha = 2$, and $\tau = 7\,\text{dB}$

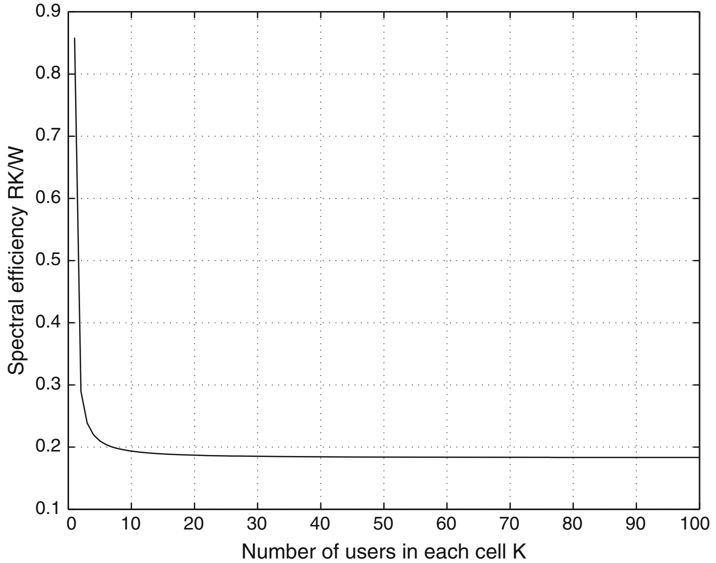


Fig. 5.29 Spectral efficiency with out-of-cell interference: $Q^{-1}(p_{\text{out}}) = 2$, $\alpha = 2$, and $\tau = 7\,\text{dB}$

Notice that in-cell interference is a term $(K-1)P$ in total interference. As K increases, the normalized interference also increases, and becomes more important than the interference averaging.

5.7.2 Hopping Method

Periodic hopping patterns consider frequency diversity and interference diversity. Frequency diversity is exploited by allocating subcarriers as spread as possible and alternate them every symbol time. Interference diversity also considers to allocate hop patterns that are as “apart” as possible from adjacent base stations. If there are N subcarriers then N can be selected as the hopping period and there can be N channels.

Construction of hopping pattern is based on two criteria: hopping requirement over all the subcarriers in each period and every channel occupies different subcarrier in each symbol. Considering these two requirements, row and column of a matrix with size $N \times N$ would contain every channel number $(1,\ldots,N)$. Such a matrix is called *Latin square*.

5.7.3 Latin Square

A Latin square is an $N \times N$ matrix with N different numbers in such a way that each symbol occurs exactly once in each row and in each column. An example is given for $N \in \{1,2,3,4,5\}$;

$$\begin{bmatrix} 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 2 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \\ 3 & 4 & 1 & 2 \\ 4 & 3 & 2 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 4 & 1 & 5 & 3 \\ 3 & 5 & 4 & 2 & 1 \\ 4 & 1 & 5 & 3 & 2 \\ 5 & 3 & 2 & 1 & 4 \end{bmatrix}$$

where the orthogonal array representation for $N = 3$ is (1,1,1), (1,2,2), (1,3,3), (2,1,2), (2,2,3), (2,3,1), (3,1,3), (3,2,1), (3,3,2). Latin square is written as a triple (r,c,s), where r is the row, c is the column, and s is the symbol for a set of N^2 triples. There is no known easily-computable formula for the number of Latin squares. Number of Latin squares can grow exceedingly quickly when N increases. For example, number of Latin squares for N from 1-to-11 are

N	No. of Latin squares
1	1
2	2
3	12
4	576
5	161280
6	812851200
7	61479419904000
8	108776032459082956800
9	5524751496156892842531225600
10	9982437658213039871725064756920320000
11	776966836171770144107444346734230682311065600000

Each base station has its own Latin square and the assignment of Latin square considers to have minimal overlap between channels of adjacent base stations. Two Latin squares are considered to be orthogonal if there is exactly one symbol/subcarrier collision for every pair of channels. This way frequency diversity and interference diversity is achieved.

5.7.4 Flash-OFDM Architecture

Flash-OFDM is suitable for sparse networks, since the number of active users is limited within a cell. Flash-OFDM considers N to be prime number to simplify the construction of the pattern and base stations need to be time and frequency synchronized. Table 5.2 shows the Flash-OFDM PHY parameters. There are 113 channels and these channels are classified into four downlink and five uplink traffic channels with different sizes as seen in Table 5.3. Scheduler performs allocation in each frame, which is 1.5-ms long. If a user is assigned to traffic channel *one* in uplink with 16QAM, then the user can transmit 7 × 14 × 4 bits.

Figure 5.30 shows the master operation of Flash-OFDM system. A user cycles in three states: ON (∼30 users), HOLD (∼130 users), and SLEEP (∼1,000 users). The users are in ON state when they are actively communicating with the base station. When they are inactive they go to SLEEP mode where they only monitor once in a while and turn off most of their power-consuming functionalities. HOLD state is basically an active state with maintaining all the synchronization functionalities in

Table 5.2 Flash-OFDM PHY parameters

Bandwidth	1.25 MHz
N	113
No. of channels	113
Chip rate	1.25 MHz
Cyclic Prefix	11% /16 chips
Max Delay Spread	11 µs
Max Doppler Spread	200 Hz
OFDM symbol length	100 µs
Frame	14 OFDM symbols
Frame length	1.4 ms
Peak downlink rate	2.7 Mbps
Peak uplink rate	750 Kbps

Table 5.3 Flash-OFDM Traffic channels

Channel No.	Downlink	Uplink
1	7	48
2	14	24
3	14	12
4	14	12
5	28	

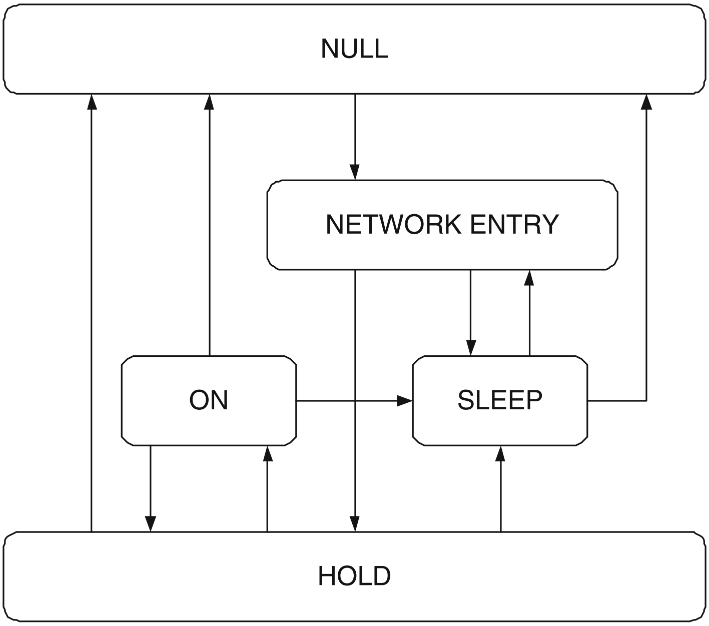


Fig. 5.30 Flash-OFDM state machine

the ON state except power control, but in HOLD state, a user is ready for transmission and waits either for packet to transmit or if there is packet then waits to be scheduled. HOLD state is suitable for packet-based applications where the traffic is typically in bursts and interleaved with large random inactive times.

5.8 Subcarrier Sharing: Embedded Modulation

Now, we introduce a different resource allocation that also incorporates constellation points in a modulation. This scheme allows a subcarrier to be used by more than one user. This scheme is known as embedded modulation and exploits the fact that a subcarrier that is high quality to a user may be high quality to multiple users and therefore that subcarrier is utilized to carry bits of multiple users.

For example, Fig. 5.31 shows the case for 64QAM embedded constellation, where three 4-QAM is stacked. Decoding is shown by arrows first, outer 4-QAM is detected, then the middle one, and finally inner 4-QAM. Each carrier carries 6 bits for 64QAM modulation and each user uses only 2 out of 6. Notice that power allocation for user in that subcarrier varies with respect to the position in stacking. User assigned for $s = 1$ would be determined similar to (5.11) for 4-QAM. For example, if users experience the same channel gain and bit error rate then the second user in the stacking $s = 2$ would have 4-fold power allocation of first user and third user in the stacking $s = 3$ would have 4-fold power allocation for second user.

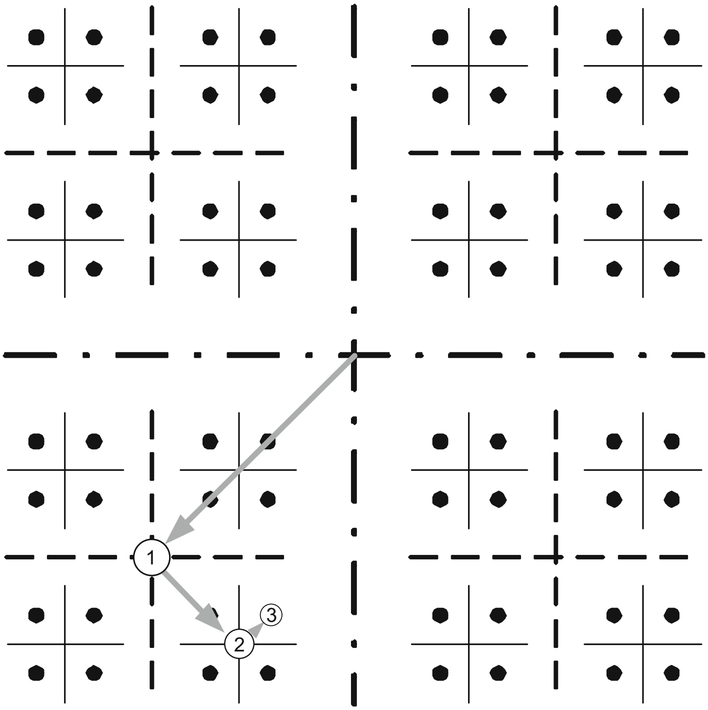


Fig. 5.31 Embedded modulation for 64QAM: Three 4-QAM modulation is embedded to address three different users

5.8.1 Optimum Solution

Optimum solution for a downlink system considers K users, N subcarriers, and S stacks within a subcarrier. Stacks are created with M-QAM symbols for $s \in \{1,2,\ldots,S\}$. Now, $\gamma_{k,n,s} = 1$ indicates user k is stacked to subcarrier n. The block model for downlink is presented in Fig. 5.32. Resource allocation problem tries to find the $\gamma = [\gamma_{k,n,s}]$ and $P_{\min} = [P_{k,n,s}]$. After allocation the created waveforms within a subcarrier are modulated and transmitted with OFDM transmission.

Problem formulation for fixed QoS where power is minimized is given below:

$$\min \sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{s=1}^{S} \gamma_{k,n,s} P_{k,n,s} \tag{5.50}$$

subject to;

$$\begin{aligned} \mathrm{QoS}_u \leq \mathrm{QoS}_o = &\textstyle\sum_{n=1}^{N} \sum_{s=1}^{S} \gamma_{k,n,s}; \quad k = 1,\ldots,K, \\ &\textstyle\sum_{k=1}^{K} \gamma_{k,n,b} \leq 1; \quad s = 1,\ldots,S; \quad n = 1,\ldots,N, \\ &\textstyle\sum_{k=1}^{K} \sum_{s=1}^{S} \gamma_{k,n,s} \leq S; \quad k = 1,\cdots,K \end{aligned} \tag{5.51}$$

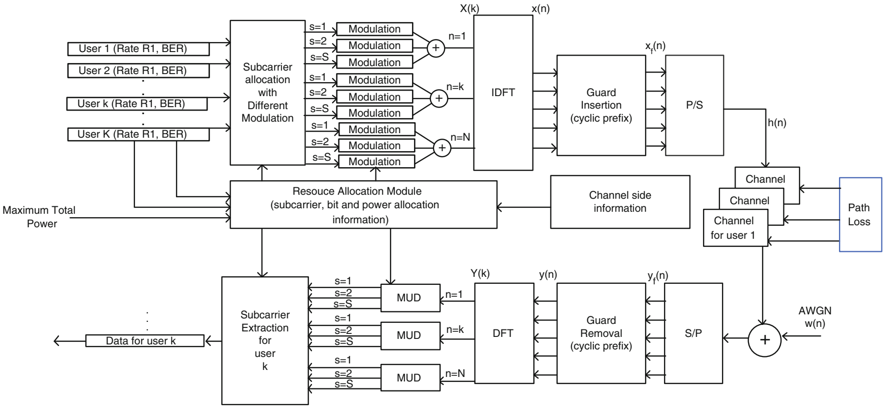


Fig. 5.32 OFDMA model for subcarrier sharing within cell

where $\gamma_{k,n,s} \in \{0,1\}$. Again, it has mutual dependency for $P_{k,n,s}$, since an allocation depends on the previous allocations in the stacking. Since positions in the higher stacking $s > 1$ has also elevation power associated with it.

5.8.2 *Iterative Solution*

Iterative solution first linearizes the system by expanding a user to w_k virtual users. Rate of a user QoS_k over bits m for minimum modulation would give the number of stacks to be allocated with minimum modulation. Total number of users (K) are now expanded to virtual users (V), where $V = \sum_k^K w_k$ and $w_k = \text{QoS}_k/m$.

Dough laying algorithm is one of the iterative scheme that considers layers one by one. In first layer ($s = 1$), allocation is done with respect to channel gain matrix and number of virtual users are restricted to $V_1 = V$. The resource allocation is similar to subcarrier allocation, which can be solved by IP, Hungarian method, or iterative scheme of Sect. 5.4.3.

For the second ($s = 2$) and third ($s = 3$) layers, the channel gain matrix is modified and allocated users in the previous layer are removed. The allocation is performed with the new channel gain matrices with $V_2 = V - N$ and $V_3 = V - 2N$ from $V_s = V - (s-1)N$.

This method can be formulated for uplink as well in which channel gains are different for each user. Embedded modulation can be enabled with preequalizer or postequalizer and for two user case; dual-signal receiver can also be considered as an option to perform the embedded modulation and demodulation.

5.9 Summary

In this chapter, we focused on scheduling algorithms of OFDMA. OFDMA is a multiple access scheme over the OFDM technology. OFDMA inherits features of TDMA, FDMA, as well as CDMA and CSMA in order to provide a flexible scheduler to perform resource allocation either based on multiuser diversity or frequency diversity or interference diversity with adaptive modulation and coding.

We first considered the problem of resource allocation in a single cell, where the objective is either minimum power subject to given fixed QoS constraints or maximum throughput subject to given fixed power. We formulated the optimum solution and then described the iterative solutions.

Second, we looked at single-frequency network in which a single frequency is used by multiple cells. We first presented the optimum solution and then looked at the heuristic solutions.

Third, we talked about Flash-OFDM system as a variant of OFDMA. We analyzed in-cell and adjacent-cell interference against CDMA and evaluated the spectrum efficiency.

Last, we presented subcarrier sharing within a cell with embedded modulation. Embedded modulation allows to stack up to S M-QAM signals in one OFDMA subcarrier. Resource allocation for such a problem is handled by layering and then allocating the resources for each layer.

References

1. Ergen, M., Coleri, S., Varaiya, P., "QoS Aware Adaptive Resource Allocation Techniques for Fair Scheduling in OFDMA Based Broadband Wireless Access Systems," *IEEE Transactions on Broadcasting,* vol. 49, 2003.
2. Khun, H. W., "The Hungarian Method for the Assignment Problem," *Naval Research Logistics Quarterly 2,* pp. 83–97, 1955.
3. Sari, H., Karam,G., Jeanclaude, I., "Transmission Techniques for Digital Terrestrial TV Broadcasting," *IEEE Communications Magazine,* vol. 33, no. 2, pp. 100–109, 1995.
4. Viswanath, P., Tse D. N. C., Laroia R., "Opportunistic Beamforming Using Dumb Antennas," *IEEE Transactions on Information Theory,* vol. 48, pp. 1277–1294, June 2002.
5. Zander, J., Kim, S.-L., *Radio Resource Management for Wireless Networks,* Artech House, Boston, 2001.
6. Pietrzyk, S., *OFDMA for Broadband Wireless Access,* Artech House, Boston, 2006.
7. Andrews, J. G., Ghosh, A., Muhammed, R., *Fundamentals of WiMAX,* Prentice Hall, New Jersey, 2007.
8. Proakis, J. G., *Digital Communications,* McGraw-Hill, New York, 1995.
9. Chua, S. G., Goldsmith, A., "Adaptive coded modulation for fading channels," *IEEE Transactions on Communication,* vol. 46, no. 5, pp. 595–602, May 1998.
10. Tse, D., Viswanath, P., *Fundamentals of Wireless Communication,* Cambridge University Press, Cambridge, 2005.
11. Jalil, R., Ergen, M., "Method and System for Managing Communication in a Frequency Division Multiple Access (FDMA) communication," United States Patent 20070297363. `www.uspto.gov.`

12. Jalil, R., Ergen, M., Mak, T., “Method and System for Spectrum Reuse in the downlink in a wireless communication network, ” United States Patent 20080139231. www.uspto.gov.
13. Bohdanowicz, A., Janssen, G. J. M., Pietrzyk, S., “Wideband indoor and outdoor multipath channel measurements at 17 GHZ,” *VTC,* vol. 4, pp. 1998–2003, 1999.
14. Goldsmith, A., Varaiya, P., “Capacity of fading channels with channel side information,” *IEEE Transactions on Information Theory,* pp. 1986–1992, November 1997.
15. Yagoobi, H., “Scalable OFDMA Physical Layer in IEEE 802.16. WirelessMAN,” *Intel Technology Journal,* vol. 8, August 2004.
16. Li. G, Liu, H., “On the optimality of the OFDMA network,” *IEEE Communications letters,* vol. 9, no. 5, pp. 438–440, May 2005.
17. Pietrzyk, S., Janssen, G. J., “Subcarrier allocation and power control for QoS provision in the presence of CCI for the downlink of cellular OFDMA systems,” *IEEE VTC,* pp. 2221–2225, April 2003.
18. Kivanc, D., Li, G., Liu, H., “Computationally efficient bandwidth allocation and power control for OFDMA,” *IEEE Transactions on Wireless Communications,* vol. 2, no. 6, pp.1150–1158, May 2004.
19. Nogueroles, R., Bossert, M., Donder, A., Zyablov, V., “Performance of a random OFDMA system for mobile communications,” *Proceedings of 1998 International Zurich Seminar on Broadband Communications,* pp. 37–48, 1998.
20. Wong, C. Y., Cheng, R. S., Letaief, K. B., Murch, R. D., “Multiuser OFDM with adaptive subcarrier, bit and power allocation,” *IEEE Journal on Selected Areas in Communications,* vol. 17, no. 10, pp. 1747–1758, October 1999.
21. Pietrzyk, S., Janssen, G. J. M., “Multiuser subcarrier allocation for QoS provision in the OFDMA systems,” *Proceedings of VTC 2002,* vol. 2, pp. 1077–1081, 2002.
22. Zhang, Y., Letaief, K. B., “Multiuser subcarrier and bit allocation along with adaptive cell selection for OFDM transmission,” *ICC,* vol. 2, pp. 861–865, 2002.
23. Rhee, W., Cioffi, J. M., “Increase in capacity of multiuser OFDM system using dynamic subchannel allocation,” *Proceedings of VTC,* vol. 2, pp. 1085–1089, 2000.
24. Wiswanath, P., Tse, D. N. C., Laroia, R., “Opportunistic beamforming using dumb antennas,” *IEEE Transactions on Information Theory,* vol. 48, no. 6, pp. 1277–1294, June 2002.
25. Knopp, R., Humblet, P., “Information capacity and power control in single cell multiuser communications,” *ICC,* vol. 1, pp. 331–335, 1995.
26. Barbarossa, S., Pompili, M., Giannakis, G. B., “Time and frequency synchronization of orthogonal frequency division multiple access systems,” *ICC,* vol. 6, pp. 1674–1678, 2001.

Chapter 6
Multiple Antenna Systems

6.1 Introduction

So far we have seen that demand for higher data rates at better quality of service is challenged by scarce usable radio resource and time-varying radio environment affected by fading and multipath.

Utilizing multiple antennas at the receiver and transmitter is widely touted as the key technique that markedly improves the data rate on longer range without consuming extra bandwidth or transmit power. This technology is also referred as multiple-input multiple-output (MIMO) communication. MIMO is now a well-mature technology; Fig. 6.1 depicts the increasing number of MIMO-related patents issued annually.

MIMO communication exploits several diversity properties often collectively referred as spatial diversity. Diversity in general has been investigated to increase the bit error probability of transmission when a channel is in deep fade. Without diversity, bit error probability decays with $\sim\frac{1}{\mathrm{SNR}}$ for high SNR as seen in Table 6.1. This basically shows that there is an unavoidable probability for a deep fade in the path that may cause error in the communication link. In this chapter, we demonstrate how one increases reliability and capacity and reduces transmission power by exploiting the spatial diversity.

Diversity techniques fundamentally exploit different paths that experience different fades. Paths can be created in time, frequency, or space to carry the same information to the receiver so that the introduced redundancy allows obtaining a more reliable signal after all paths combined in the receiver. For instance, for K paths that transmit the same symbol, an error occurs when[1] $\|\mathbf{h}\|^2 = \sum_{l=1}^{K} |h_k|^2$ is smaller than $\frac{1}{\mathrm{SNR}}$ if h_k is independent Rayleigh channel gain. With $\frac{1}{\mathrm{SNR}}$ probability $|h_k|^2$ is less than $\frac{1}{\mathrm{SNR}}$, and consequently, error probability is around $\frac{\sim 1}{\mathrm{SNR}^K}$, where K is diversity gain at high SNR. From Fig. 6.2, we can see that K determines how the slope of average probability error changes as a function of average SNR.

[1] The Frobenius norm.

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_6, © Springer Science+Business Media LLC 2009

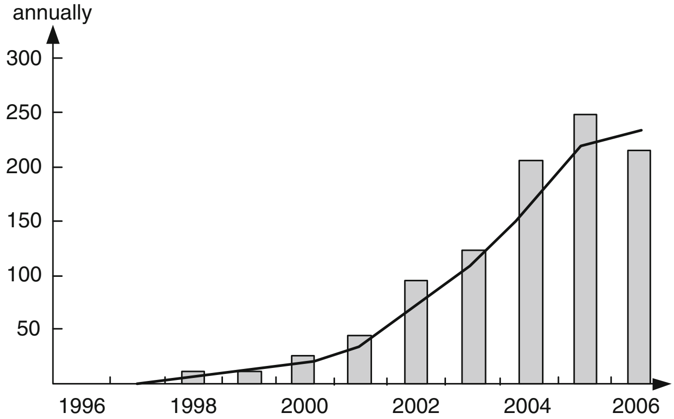


Fig. 6.1 MIMO patent applications per year. (Source: Marvedis)

Table 6.1 Modulation performance of coherent schemes under Rayleigh Fading. (Source: Fundamentals of Wireless Communications by Tse and Viswanath)

Scheme	Bit error prob	Data rate (bits/s/Hz)
Coherent BPSK	$\frac{1}{4\mathrm{SNR}}$	1
Coherent QPSK	$\frac{1}{2\mathrm{SNR}}$	2
Coherent 4-PAM	$\frac{5}{4\mathrm{SNR}}$	2
Coherent 4-PAM	$\frac{5}{4\mathrm{SNR}}$	2
Coherent 16QAM	$\frac{5}{2\mathrm{SNR}}$	4

Noncoherent modulation also shows similar performance degradation with $K\frac{1}{\mathrm{SNR}}$, where $K \in \{1/2, 1\}$ for Diff. BPSK and Diff. QPSK respectively with the same data rate

Repetition coding is a simple *time diversity* technique and the main idea is repeating the same signal over different coherence time periods. Alternatively, multiple versions of the same signal may be transmitted at different time instants. OFDM is one of the *frequency diversity* technique where the signal can be spread over a wider frequency where coherence bandwidth is smaller than the channel bandwidth. The use of time or frequency diversity reduces data rate because of the redundancy introduced in the system. In general, Fig. 6.2 shows the coding gain vs. the diversity gain.

In spatial diversity, antennas are sufficiently separated apart to create independent paths. Experimental evidence demonstrates that the number of multipath is enormous. This multipath phenomenon is leveraged to augment the rate of data transmission or to achieve more robust transmission.

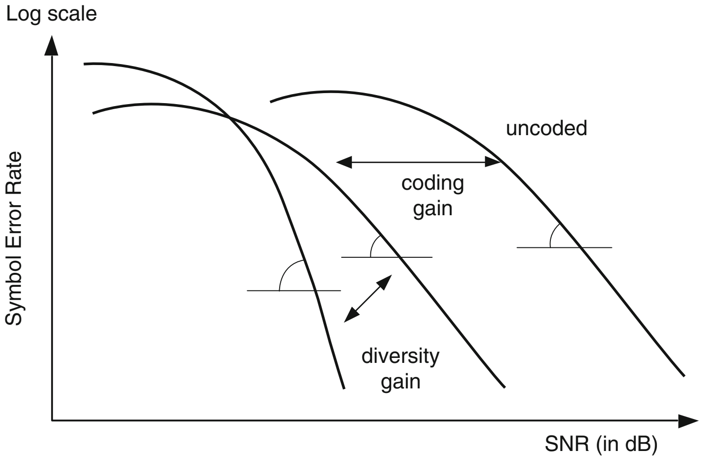


Fig. 6.2 Diversity gain. Diversity gain affects the slope of the curve. Higher diversity gain results in sharp drop. Coding gain affects horizontal shift to origin. Greater the coding gain, more shift is observed

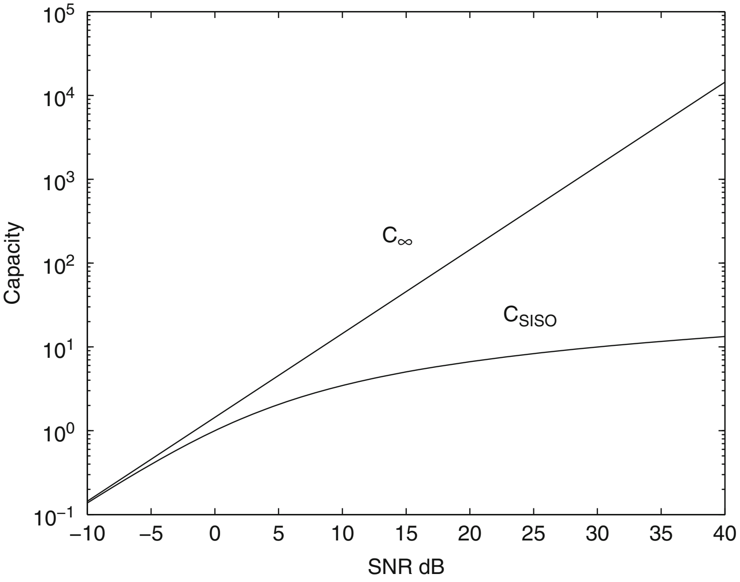


Fig. 6.3 Multiple antennas at the transmitter and receiver compared with single-input–single-output system

The theoretical capacity bound as depicted in Fig. 6.3 states that as there are increase in the number of transmitting and receiving antennas, the capacity (C) of the wireless channel tends to

$$C_\infty = \frac{\text{SNR}}{ln2}\,\text{bps/Hz}, \tag{6.1}$$

where we can compare this to the throughput (C_{SISO}) of the perfect single-input–single-output (SISO) system, which is given by Shannon's capacity formula per unit bandwidth as follows:

$$C_{\text{SISO}} = \log(1 + \text{SNR})\,\text{bps/Hz}. \tag{6.2}$$

One can see the gain to achieve by increasing antennas at the transmitter and receiver.

The emphasis of this chapter is on spatial diversity techniques that utilizes multiple antennas. In this chapter, we first describe the spatial diversity with its components. We then consider the single-input–multiple-output (SIMO) system and describe the concept of maximal ratio combining (MRC) as a way to exploit receive diversity. Next, we follow the discussion by introducing the multiple-input–single-output (MISO) system and describe how to exploit transmit diversity with space-time coding. Finally, the MIMO system is introduced, which shows both transmit and receive diversity as well as spatial multiplexing, which are the remarkable properties of MIMO that allows to increase capacity. We conclude the chapter with MIMO's integration to IEEE 802.11n technology. MIMO and OFDMA in WiMAX, LTE, and beyond is discussed in their respective chapters in detail.

6.2 Spatial Diversity

Spatial diversity utilizes multiple antennas at the transmitter and receiver with various configurations as illustrated in Fig. 6.4. Antennas are spatially separated with a fixed distance and the distance of separation depends on the carrier frequency and scattering environment.

Let us first consider a transmit antenna and two or more receive antennas; one can create two or more paths and combine the signals in the receiver. This exploits receiver diversity and the scheme is known as single-input–multi-output (SIMO) system. SIMO architecture collects more energy at the receiver to improve the SNR as compared with a SISO system.

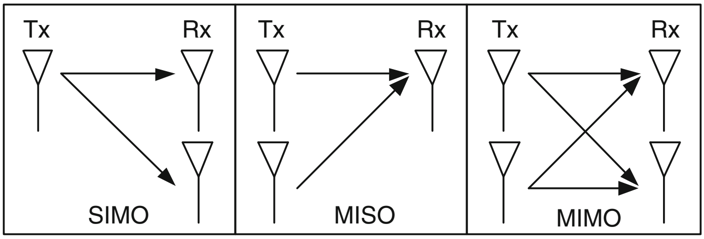


Fig. 6.4 Spatial diversity techniques

Now, consider two or more transmit antennas and one receive antenna. One can create two or more paths easily in a fading environment. This is called transmit diversity and it is known as multi-input–single-output (MISO) system. MISO achieves the same diversity gain as SIMO, nevertheless MISO can implement a different type of coding called *space-time coding*. Space-time transmitter encodes and modulates the information bits in space and time.

MIMO exploits multiple antennas at transmitter and receiver and inherits the benefits of SIMO and MISO systems and introduces more. MIMO basically uses the degrees of freedom of the channel to achieve greater performance as compared with SIMO and MISO. Also, MIMO opens up multiple independent data paths over a link. Space-time transmitter sends the signal, and space-time receiver processes the signals received on each of the receive antennas according to space-time transmitter's signaling strategy.

Before delving into more detail, further discussion about types of spatial gain is necessary:

- *Array gain* is defined as the average increase in SNR, typically linear with number of antennas irrespective of channel correlation. In other words, if there are M antennas, there is M-fold increase in the SNR and probability of error is linear with $\simeq(M.\mathrm{SNR})^{-1}$. The signals are coherently combined to improve the signal strength. Note that even if the channels are correlated (line-of-sight) the array gain is still present and SNR increases linearly. Also note that linear increase in SNR also increases the capacity according to Shannon's formula $C = \log_2(1+\mathrm{SNR})\,\mathrm{bps/Hz}$.
- *Diversity gain* is defined as the reduction in the error probability due to multiple independent (uncorrelated) paths created between the transmitter and receiver. In other words, if there are K transmit, M receive antennas, the diversity order is $K.M$, and the error probability improves proportionally with $\simeq \mathrm{SNR}^{-K.M}$. If the medium is line-of-sight channel, there is no diversity gain.
- *Multiplexing gain* is defined as the increase in the data rate, since independent paths between multiple transmitters and multiple receivers may be utilized to send independent data streams. In other words, if there are K (>1) transmit antennas and M (>1) receive antennas, the increase in the data rate is $\min(K,M)$-fold.

6.3 Basics of MIMO

Now, we introduce a general MIMO system. We follow the notation introduced here throughout the chapter and derive SIMO and MISO from this generalized introduction.

6.3.1 MIMO Channel

Let us consider a MIMO channel (Fig. 6.5) for reference in this chapter with K transmit and M receive antennas (Note if $K = 1$, it is SIMO, if $M = 1$, it is MISO, and if $K = M = 1$, it is a SISO system). There are $K \times M$ paths and each path has a channel response denoted by h_{ij}, which is between ith receiver and jth transmitter. The MIMO channel (**H**) is shown below,

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1K} \\ h_{21} & h_{22} & \cdots & h_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M1} & h_{M2} & \cdots & h_{MK} \end{bmatrix} \tag{6.3}$$

and according to this **H**, if the transmitted signal is,

$$\mathbf{x} = [x_1, x_2, \ldots, x_K]^{\mathrm{T}}, \tag{6.4}$$

the signal received at the receive antenna is as follows:

$$\mathbf{y}_{M\times 1} = \mathbf{H}_{M\times K}\mathbf{x}_{K\times 1} + \mathbf{n}_{M\times 1}, \tag{6.5}$$

where **n** is the noise vector consisting complex-gaussian elements with zero mean and variance σ_n^2. Sufficient antenna separation (typically half the carrier wavelength

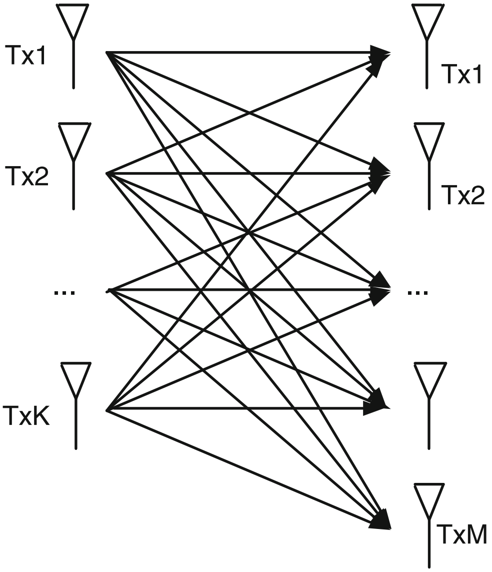


Fig. 6.5 MIMO channel

($\frac{\lambda}{2}$)) makes elements of $\mathbf{H}$ independent, zero-mean, complex Gaussian random variables (Rayleigh fading). However, at a given time, $\mathbf{H}$ varies over frequency and time depending on multipath and Doppler spread respectively.

6.3.2 Decoding

Let us look at first decoding techniques attributed in this chapter:

- *Maximum-likelihood (ML) decoder* is an optimum decoder that finds $\hat{x}$ that minimizes the distance $\hat{\mathbf{x}} = \arg\min \parallel \mathbf{y} - \mathbf{H}\hat{\mathbf{x}} \parallel$. Notice that the search to find right input is computationally complex as it requires searching among m^K inputs, where m is the modulation (e.g, $\mathbf{m} = \mathbf{16}$ for 16QAM) and K is number of transmit antennas. ML decoder is used when channel side information is absent in the transmitter. If it is known, the gain from channel information is minimal.
- *Zero-forcing (ZF) decoder* is a linear decoder that recovers the transmitted $\mathbf{x}$ by multiplying the received signal with $\mathbf{G} = \mathbf{H}^{-1}$ as $\hat{\mathbf{x}} = \mathbf{G}\mathbf{H}\hat{\mathbf{x}} + \mathbf{H}^{-1}\mathbf{n}$. Notice that interference from other antennas are removed, but inverse of $\mathbf{H}$ might boost up the noise as bad subchannels that have lower eigenvalues are inverted. This can easily amplify the noise.
- *Minimum-mean-square-error (MMSE) decoder* on the other hand balances the noise enhancements and interference from other antennas by minimizing the distortion in $\mathbf{G}$ by

$$\mathbf{G} = \arg\min_{\mathbf{G}} E(\parallel \mathbf{G}\mathbf{y} - \mathbf{x} \parallel^2), \tag{6.6}$$

  where $\mathbf{G}$ is $(\mathbf{H}^H\mathbf{H} + \frac{1}{\mathrm{SNR}}\mathbf{I})^{-1}\mathbf{H}^H$. This prevents worst eigenvalues being inverted at low SNR and converges to ZL decoding at high SNR.

6.3.3 Channel Estimation

Now let us look at channel estimation for MIMO system as decoding requires an estimate of $\mathbf{H}$ at the receiver and some of them require channel side information (CSI) at the transmitter. Training-based channel estimation is suitable for MIMO systems and chosen also in the standards.

Training-based channel estimation requires transmitting known symbols such as preambles and pilots. Typically, in IEEE 802.16 and IEEE 802.11, preambles are used for synchronization and channel estimation, and pilots are used for fine tuning synchronization and channel estimation if channel is time-varying.

Note that the received signal is superposition of K transmit antennas. Thus, training signals need to be orthogonal in each transmit antenna so as to be transmitted without interference. Hence, when a pilot is transmitted in a subcarrier, the other entities are kept silent for that subcarrier for the same symbol. This of course gives partial information about the channel; the whole channel response is obtained with interpolation. LS or MMSE channel estimation (see Chap. 4) can be used in frequency domain to estimate the channel.

6.3.4 Channel Feedback

Channel side information is required in the transmitter for some of the schemes introduced below to perform precoding. This yields better performance and enables complex signal processing techniques. In TDD systems, channel is considered reciprocal, where downlink channel is obtained from uplink channel. For instance, in WiMAX-e (IEEE 802.16e), there are sounding zones in the uplink to obtain the CSI at the transmitter for downlink.

Another technique is feedback channel, required in FDD and optional in TDD systems. Feedback channel transmits compressed channel information, which is necessary for precoding in the transmitter. Typically, rather than sending the channel information, at the receiver, required precoding technique is computed and sent back to the transmitter. Precoder is constrained to be one of ξ distinct matrices and requires $B = \log_2 \xi$ bits of feedback. Of course, constrained precoder deviates from the optimal one.

Jacobi rotation is a nondifferential feedback method and sends the Jacobi rotation matrix $\mathbf{J}$ or its index as a feedback. Jacobi matrix diagonalizes the channel correlation matrix $\mathbf{R} = \hat{\mathbf{H}}\hat{\mathbf{H}}^H$, where $\hat{\mathbf{H}}$ is the estimate of channel response matrix and can be decomposed using SVD by

$$\hat{\mathbf{H}} = UDV^H, \tag{6.7}$$

where U and V are the unitary matrices, i.e., $U^H U = I$ and $V^H V = I$. D is a diagonal matrix with singular values in the diagonal.

$\mathbf{R}$ can be rewritten as $\mathbf{R} = \mathbf{V}\mathbf{D}^2\mathbf{V}^H$ with substitution of (6.7). Alternatively Jacobi rotation is used to perform the diagonalization of $\mathbf{R}$ as follows

$$D^2 = \mathbf{J}^H\mathbf{R}\mathbf{J} \tag{6.8}$$

and $\mathbf{J}$ is used as the precoding matrix. For 2×2 MIMO, $\mathbf{J}$ is given as

$$J(\theta,\phi) = \begin{bmatrix} \cos(\theta)e^{j\phi} & \sin(\theta)e^{j\phi} \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}, \tag{6.9}$$

where θ and ϕ can be obtained from the following:

$$\begin{array}{l} \tan(\theta)^2 + \frac{(r_{22}-r_{11})}{|r_{12}|}\tan(\theta) - 1 = 0 \\ e^{j\phi} = \frac{r_{12}}{|r_{12}|} \end{array}. \tag{6.10}$$

Also, differential feedback with Jacobi transformation is possible if Jacobi rotation for nth instance is

$$\mathbf{J}(n)^H\mathbf{R}(n)\mathbf{J}(n) = \mathbf{D}^2 \tag{6.11}$$

and for the next feedback instance $(n+1)$, the Jacobi rotation is given by

$$\mathbf{J}(n)^H\mathbf{R}(n+1)\mathbf{J}(n) = \bar{\mathbf{D}}^2, \tag{6.12}$$

where $\bar{\mathbf{D}}$ is not diagonal; $\mathbf{J}(n+1)$ for the $n+1$ instance is found with differential precoding matrix $\mathbf{J}$ as follows

$$\mathbf{J}(n+1) = \mathbf{J}(n)\mathbf{J}. \tag{6.13}$$

Depending on the channel condition, differential feedback may be selected for slow varying channels, and nondifferential feedback is selected for high speed channels. Combination of these two is possible with different codebooks and resetting at some period is necessary to stop error accumulation. The codeword in the codebook that might represent the $\mathbf{J}$ is selected according to minimum distance between the parameters of $\mathbf{J}$ and codeword vector according to Lloyd algorithm. Feedback rate depends on the coherence time T_C, feedback bits B, and periodicity p. Hence, the required data rate is $\frac{B\times p}{T_C}$ bps.

Feedback processing and delay for channel side information impacts the performance. The receiver often quantizes and also averages across all subcarriers to reduce the overhead of the precoding matrix and during this process some information may be lost. Also, if channel varies very fast, the degradation further increases.

6.4 SIMO

We start with introducing SIMO first to analyze array gain and diversity. Consider a SIMO system with one transmit and M receive antennas. M antennas in the receiver receive $y_1, y_2, \dots, y_M$ signals. When signals are transmitted, the channel determines the propagation conditions such as amplitudes and phases. Each path experiences a phase distortion ϕ_i, and this distortion is compensated in the receiver with a multiplication by $\beta_i = c_i e^{-j\phi_i}$ for some real-valued c_i. This procedure is called co-phasing, and it is necessary to remove the fading effect in the signal.

When signals are coherently combined in the receiver, the resultant power of the signal is enhanced in quality because of array gain. The average increase in the signal power is M-fold and proportional to the number of receive antennas.

In addition to array gain, SIMO systems benefit from *receiver diversity* as well. Diversity gain brings several realizations of the transmitted signal to the receiver and receiver applies one of the several combining techniques so that the resultant signal exhibits reduced fade and amplitude variability.

6.4.1 Combining Techniques

There are many methods for combining the diversity branches at the receiver; the most widely used combining techniques ranging from low to high complexity designs are described in the next section.

6.4.1.1 SDC: Selection Diversity Combining

SDC is a selection technique to sample the each branch and selecting the best branch that maximizes signal to noise ratio of the receiver.

The requirement is that the antennas should be separated or operated in different phases or both to have an independent channel. The spatial correlation is approximated by the zero-order Bessel function given by the equation $\rho = J_0^2(2\pi d/\lambda)$, and the distance beyond 1/3 of wavelength is considered adequate to achieve an independent channel.

Let us denote the instantaneous received symbol energy-to-noise ratio (E_b/N_o) by τ_i, $i = 1,\ldots,M$ on the ith diversity branch. τ_i has Rayleigh fading with the following pdf:

$$f_{\tau_i}(x) = \frac{1}{\bar{\tau}} e^{-x/\bar{\tau}}, \tag{6.14}$$

where $\bar{\tau}$ is the average received branch E_b/N_o.

With selective combining, the output of the receiver is the one with highest SNR branch:

$$\tau_{\text{SDC}} = \max\{\tau_1,\ldots,\tau_M\}, \tag{6.15}$$

where the cumulative distribution function (CDF) is given by

$$F_\tau(x) = \Pr[\tau_1 \le x,\ldots,\tau_M \le x] = [1 - e^{-x/\tau}]^M \tag{6.16}$$

if the average SNR for all of the branches are the same. Also, the SNR for SDC operation is found as

$$\tau_{\text{SDC}} = \bar{\tau} \sum_{i=1}^{M} \frac{1}{i}, \tag{6.17}$$

where we can see that the gain from the additional antenna diminishes with more antennas.

Figure 6.6 shows the CDF of output of selective combiner. Gain increases with $10\log(n)$, where n is the number of antenna array elements, and doubling the antenna doubles the gain.

SDC operation should operate within the coherence time so that constant phase is preserved both temporally and spatially. SDC is impractical for systems that use continuous transmission because it requires monitoring each diversity branch.

In brief, the SDC scheme exhibits no array gain, since only one antenna is used. On the other hand, it achieves stellar *receiver* diversity gain with spatial or polarization diversity of the receiver. SDC can be modified to threshold combining as well

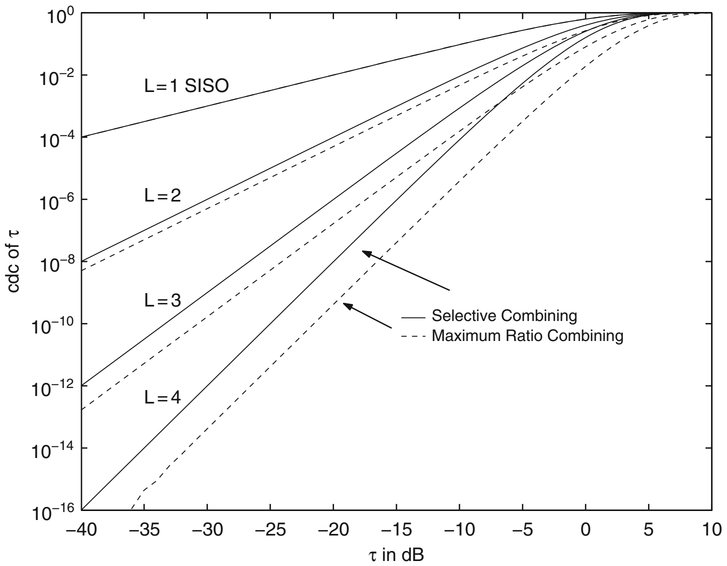


Fig. 6.6 Selective Combining and Maximum Ratio Combining where $\bar{\tau} = 1$. Notice the largest diversity gain obtained with two branches as compared to SISO since increase in L results more gains but marginal

in which a branch is selected if the branch SNR is above a threshold and altered to the next if it goes below the threshold. Switch-and-stay combining (SSC) is a simple version for two antennas, where switch happens if branch SNR is lower than the threshold.

6.4.1.2 EGC: Equal Gain Combining

The EGC method rather than selecting the best antenna combines power of all antennas after cophasing with a parameter $e^{-\phi_i}$. This is to cophase independent branches that are set to unity before combining.

The SNR for EGC is given by

$$\tau_{\mathrm{EGC}} = \bar{\tau}\frac{|\sum_{i=1}^{M}\beta_i h_i|^2}{\sum_{i=1}^{M}|\beta_i|^2}, \tag{6.18}$$

where β_i is set to $e^{-\phi_i}$ and $c_i = 1$ in order to perform only co-phasing. As a result, the optimum SNR for EGC is

$$\tau_{\mathrm{EGC}} = \bar{\tau}\frac{|\sum_{i=1}^{M} h_i|^2}{M}. \tag{6.19}$$

The EGC utilizes *receiver* diversity gain and also array gain as compared with the SDC, since averaging the received branches may constructively increase the signal strength.

6.4.1.3 MRC: Maximum Ratio Combining

The MRC is similar to EGC except that the algorithm tries to optimally adjust both phase and gain of each element prior to combining. Processing can be done in analog or digital domain, and digital domain processing is better when compensating the frequency selective channel characteristics. MRC achieves highest antenna diversity gain as compared with others.

With maximum ratio combining, the branches are combined after weighted by a complex fading gains $c_i e^{-j\phi_i}$. c_i is selected proportional to the branch SNR $|c_i| = |h_i|$ in order to maximize the overall SNR. Hence, SNR for MRC is given by

$$\tau_{\mathrm{MRC}} = \bar{\tau} \sum_{i=1}^{M} |h_i|^2, \tag{6.20}$$

and output of MRC combiner has the following CDF:

$$F_{\tau}(x) = 1 - \sum_{i=0}^{M-1} \frac{1}{i!} (\frac{x}{\bar{\tau}})^i e^{-x/\bar{\tau}}, \tag{6.21}$$

and upper bound for probability error is given by the following for Rayleigh fading channel:

$$\Pr\{e\} \leq \frac{1}{(1+\frac{\tau}{2})^M}, \tag{6.22}$$

where M is also termed diversity gain. When $M = 1$, the equation converges to SISO in Rayleigh fading. And, compare this with the error probability in AWGN for SISO given by

$$\Pr\{e\} \leq e^{-\tau/2}. \tag{6.23}$$

Figure 6.6 shows that MRC performs better than SDC in terms of probability outage. Table 6.2 also compares the combining techniques for four antenna systems. The gains of analog MRC are 1 dB and 2 dB over EGC and SDC respectively.

6.4.1.4 SIMO Capacity

The Shannon capacity of the SISO system represented with C is as follows:

$$C_{\mathrm{SISO}} \approx \log_2(1 + \bar{\tau} h^2)\, \mathrm{bps/Hz}, \tag{6.24}$$

where $\bar{\tau} = E_b/N_o$ is the average signal-to-noise ratio.

Table 6.2 Combining techniques

Combining technique (4 antenna branches)	Antenna gain with SUI3, SUI4 model with 100 µs Rayleigh delay spread (dB)	Complexity	Gain
SDC	8	Low	Diversity gain
Analog EGC	9	Mid	Diversity and array gain
Analog MRC	10	High	Diversity and array gain
Digital MRC	14	High	Diversity and array gain

Table 6.3 Diversity order

System	Array gain	Diversity order	Multiplexing gain
SISO	1	1	
SIMO with CSI at the receiver (MRC 1×M)	M	M	
MISO with CSI at the receiver (Alamouti 2×2)	1	K	
MISO with CSI at the transmitter and receiver (Transmit beamforming K×1)	K	K	
MIMO with CSI at the receiver (Alamouti-based 2×2)	M	KM	
MIMO with CSI at the transmitter and receiver (beamforming $K \times M$)	$\leq KM$	KM	
MIMO with CSI at the transmitter and receiver (spatial multiplexing $K \times M$)			$\min(K,M)$

For SIMO system, with M receive antennas, we need to consider the increase in SNR to determine the capacity. SNR is the sum off all paths, where each path experiences a channel gain h_i^2. Therefore, capacity is given by

$$C_{\mathrm{SIMO}} = \log_2\left(1 + \bar{\tau}\sum_{i=1}^{M} h_i^2\right) \text{bps/Hz} \tag{6.25}$$

for deterministic Gaussian channel.

Intuitively, if we take $\bar{\tau}$ as basis for one path, combining M path produces M^2 increase in the signal power. Same way, noise power would increase $M{-}fold$ (see Table 6.3) and hence overall increase will be $\frac{M^2}{M}\bar{\tau}$ and C_{SIMO} is

$$C_{\mathrm{SIMO}} \approx \log_2(1 + M.\bar{\tau})\,\text{bps/Hz}. \tag{6.26}$$

6.5 MISO

Sometimes multiple antennas at the transmitter is more practical considering the small form-factor of handheld devices: multiple antennas may not be separated sufficiently far apart to achieve independent fading. Can we achieve a similar diversity and array gain as in MRC if multiple antennas are only at the transmitter? The answer is yes. Increasing the SNR and removing some fading with diversity can be achieved with *transmit* diversity, which is available in MISO systems.

In MISO systems, which has K transmit antennas, array gain is proportional to the number of transmit antennas. Previously, we have seen that receive diversity requires channel information at the receiver. For transmit diversity, there are two types of transmit diversity techniques: transmit diversity with CSI at the transmitter and transmit diversity without CSI at the transmitter.

6.5.1 Transmit Diversity with CSI

Transmitter is aware of the channel condition that is sent by the receiver via feedback channel as seen in Fig. 6.7. The channel side information (CSI) – amplitude and phase $h_i = \rho_i e^{-j\phi_i}$ – can be measured at the receiver using pilot techniques or with channel reciprocity in TDD systems.

Similar analysis for SDC and MRC can be done for transmit diversity (aka transmit beamforming) as well. In SDC, the receiver transmits back the selected antenna index, and the selected antenna operates with full power during transmission. In MRC, amplitude and phase of each branch is sent to the transmitter. The branch weights in the transmitter are set to the received SNR. The resultant SNR, which is

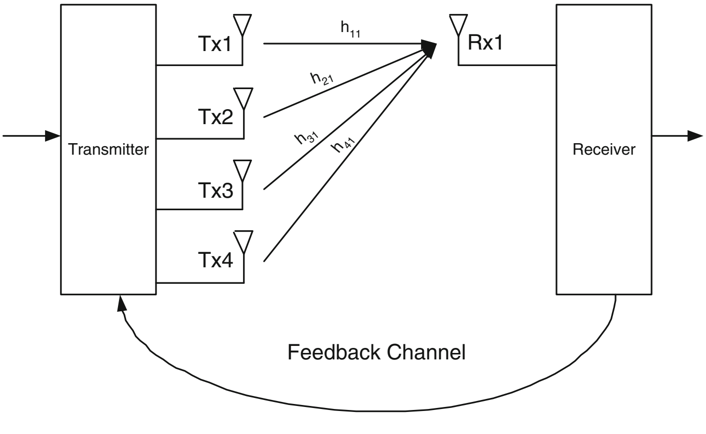


Fig. 6.7 Transmit diversity with channel side information; in SDC, the receiver sends back the antenna index; in EGC, the receiver sends back the phase of each branch; in MRC, the receiver sends back the amplitude and phase of each branch

the sum of all branches, shows K-fold increase in SNR over a single antenna with full power. Notice that branch gains are set in a way that the total transmitted power does not exceed the transmit power in SISO case, which means that the summation of weights must be one. Therefore, the resultant signal is

$$y = \sum_{i=1}^{K} \beta_i h_i x, \tag{6.27}$$

where $\beta_i = c_i e^{-j\phi_i}$, $h_i = \rho_i e^{-j\phi_i}$, and x is the transmitted signal with E_b energy. Branch weights are set in order to maximize the received SNR. Hence,

$$c_i = \frac{\rho_i}{\parallel H \parallel} \tag{6.28}$$

and the resultant SNR becomes

$$\tau_{\text{MISO}} = \bar{\tau} \sum_{i=1}^{K} |h_i|^2, \tag{6.29}$$

where the probability error is bounded by

$$\Pr\{\mathrm{e}\} \leq \frac{1}{(1+\frac{\tau}{2})^K}, \tag{6.30}$$

which is the same as in SIMO MRC.

6.5.2 Transmit Diversity Without CSI (Alamouti Scheme)

If the transmitter is not aware of the channel gain $\rho_i e^{-j\phi_i}$, obtaining a transmit diversity requires adroit coding. One of the famous system introduced is the Alamouti scheme, which considers two transmit antennas that operates with half of single transmitter energy. The subtle idea is as follows: In the first symbol at time t_0, from transmitter$_1$, the signal x_1 is transmitted, and from transmitter$_2$, the signal x_2 is transmitted. They experience h_1 and h_2 channels and $n(t_0)$ noise respectively. As a result, the received signal is $y(t_0) = x_1 h_1 + x_2 h_2 + n_1(t_0)$. In the second symbol, at time t_1, now transmitter$_1$ sends $-x_2^*$ and transmitter$_2$ sends x_1^*. Now, the received signal in the second symbol is $y(t_1) = -x_2^* h_1 + x_1^* h_2 + n(t_1)$.

Notice that assumptions here are constant channel gains over two symbol periods $h_1(t_0) = h_1(t_1)$ and $h_2(t_0) = h_2(t_1)$ and $n(.)$ is a sample Gaussian noise.

The receiver operates over those two consecutive symbols that form a block. Now, if we interpret those two symbols together, the received signal is a vector in the form of $\mathbf{y} = [y_1 y_2^*]^{\mathrm{T}}$, $\mathbf{x} = [x_1 x_2]^{\mathrm{T}}$, and $\mathbf{n} = [n_1 n_2]^{\mathrm{T}}$;

$$\mathbf{y} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2^* \end{bmatrix} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{6.31}$$

where **H** is

$$\mathbf{H} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix}. \tag{6.32}$$

We can manipulate **y** with $\mathbf{H}^H$

$$\mathbf{H}^H = \begin{bmatrix} h_1^* & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \tag{6.33}$$

in order to achieve

$$\begin{aligned} \zeta_1 &= (|h_1^2| + |h_2^2|)x_1 + h_1^* n(t_0) + h_2 n^*(t_1) \\ \zeta_2 &= (|h_1^2| + |h_2^2|)x_2 + h_2^* n(t_0) - h_1 n^*(t_1), \end{aligned} \tag{6.34}$$

where for ζ_i, the resulting SNR can be formed by

$$\tau_i = \frac{(|h_1^2| + |h_2^2|)E_b}{2N_o}. \tag{6.35}$$

Notice that the array gain is 1, which is half of the array gain in 2 × 1 MISO MRC and 1 × 2 SIMO MRC. This result is due to the reduced transmission power; remember we set the transmission power to one half in order to keep the total energy as E_b. This means that 1 × 2 MRC and 2 × 1 MRC outperform Alamouti's performance by 3 dB. However, the Alamouti scheme achieves the same diversity gain as both MRC schemes, which is 2.

In general, the upper bound of probability error for Alamouti scheme is given by

$$\Pr\{e\} \le \frac{1}{(1 + \frac{\tau_{\text{Alamouti}}}{4})^2}, \tag{6.36}$$

where

$$\tau_{\text{Alamouti}} = \frac{\sum_{i=1}^{2} |h_i|^2 E_b}{2N_o}. \tag{6.37}$$

Notice that the channel side information has increased the SNR by up to a factor of 2. The Alamouti scheme is generalized in space-time block coding (STBC) for more than two antennas.

6.5.3 MISO Capacity

In MISO, system having K transmit antennas, total transmit power is divided into K branches. The resultant SNR is combined over the air now in contrast to SIMO where they are combined in the receiver. The capacity C_{MISO} is

$$C_{\text{MISO}} = \log_2 \left(1 + \frac{\bar{\tau}}{K} \sum_{i=1}^{K} h_i^2 \right) \text{bps/Hz} \tag{6.38}$$

for deterministic Gaussian channel. Notice that MISO does not have array gain at the receiver, which makes SIMO capacity increase more rapidly than MISO capacity with number of antennas.

6.6 MIMO

MIMO systems utilize transmit and receive diversity together. A $K \times M$ MIMO system consists of $K \times M$ SISO links, and the diversity gain is equivalent to that of an MRC system with $K.M$ antennas. As a result, diversity gain scales linearly with the product of the number of receive and transmit antennas.

In MIMO system, array gain depends on the number of transmit and receive antennas and there is a $10\log_{10} K$ dB penalty in SNR since total transmit energy is divided into K. On the other hand, there is a K-fold increase if $M \geq K$, since all transmitted power is collected.

6.6.1 MIMO Beamforming – Eigenbeamforming

Consider the same signal x is transmitted over all antennas with different gains v_i, where $\| \mathbf{v} \| = 1$. In the receiver side, each branch is multiplied with u_i^* ($\| \mathbf{u} \| = 1$) as seen in Fig. 6.8. The resulting output is then

$$\mathbf{y} = \mathbf{u}^H \mathbf{H} \mathbf{v} x + \mathbf{u}^H \mathbf{n}. \tag{6.39}$$

The optimum selection of **u** and **v** at the transmitter and receiver maximizes the SNR. Hence, maximum SNR is bounded by

$$\frac{1}{\min(K,M)} \bar{\tau} ||\mathbf{H}||^2 \leq \tau_{\text{MIMO}} \leq \bar{\tau} ||\mathbf{H}||^2, \tag{6.40}$$

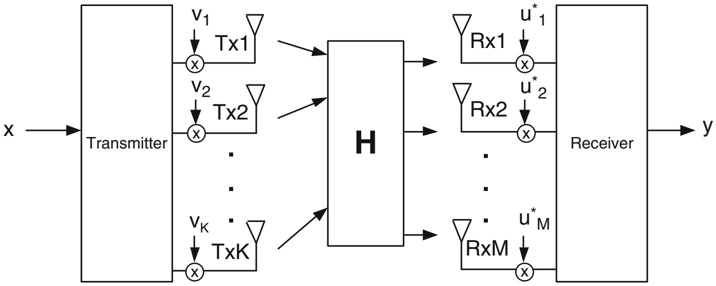


Fig. 6.8 Beamforming with MIMO system

where the error probability is upper bounded by

$$\Pr\{e\} \leq \frac{1}{\left(1 + \frac{\bar{\tau}}{2\min(K,M)}\right)^{KM}}. \tag{6.41}$$

From SVD, signal may concentrate on the highest eigenvalue of $\mathbf{H}$. Therefore, the received SNR equals to $\sigma_{\max}^2 \bar{\tau}$ and the capacity becomes $C = \log_2(1 + \sigma_{\max}^2 \bar{\tau})$.

As you note, there is no physical directionality involved in this process. The required element is channel matrix in the transmitter, and the created beam does not physically direct to a location. Basically, this technique identifies the strongest eigenvalue of the channel and mathematically steers the signals toward that eigenchannel. There is also another type of beamforming technique that manipulates the omni-directional transmission into a physically directional beam. It can be achieved by engineering the phase and amplitude of each omni-directional wave in a way that in one direction waves add up and in other direction signals cancel each other. As a result, main lobe is powerful as compared with side lobes in the wave.

Figure 6.9 compares the error probability bound for SISO, SIMO, MISO, and MIMO up to two transmit and receive antennas. From the figure, we can first see the detrimental effect of Rayleigh fading against AWGN in SISO. Second, note that SIMO MRC and MISO MRC have the same error probability bound. Third, horizontal shift from 2×1 Alamouti scheme to MISO MRC (SIMO MRC in the figure) indicates the array gain. Last, MIMO 2×2 beamforming in the figure considered CSI at the transmitter and receiver. But with Alamouti-based scheme, 2×2 and 4×2 can be constructed by only utilizing CSI at the receiver.

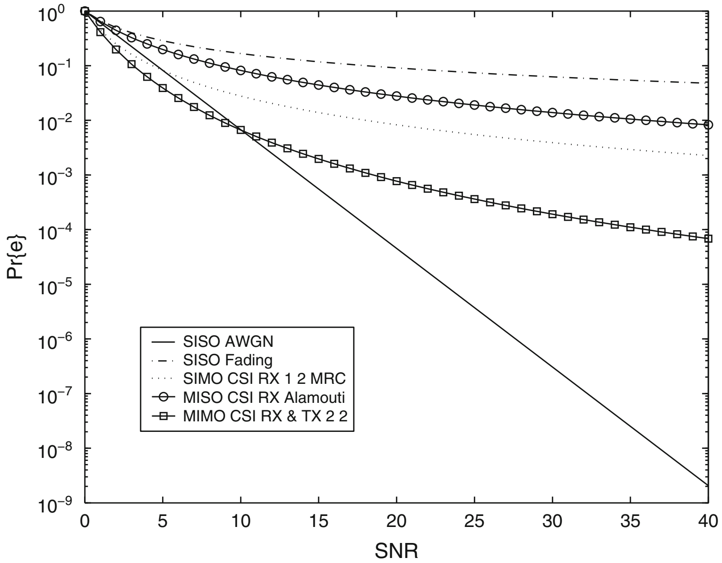


Fig. 6.9 Bound for error probability

6.6.2 2×2 MIMO – Alamouti Based

The 2×1 Alamouti scheme is configured for 2×2 MIMO over two symbol periods as depicted in Fig. 6.10. At time t_0, the receiver$_1$ and the receiver$_2$ get $y_1(t_0)$ and $y_2(t_0)$. Similarly at time t_1, the receiver$_1$ and the receiver$_2$ get $y_1(t_1)$ and $y_2(t_1)$. Hence, the received signal becomes

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} y_1(t_0) \\ y_1^*(t_1) \\ y_2(t_0) \\ y_2^*(t_1) \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} \mathbf{n}_1 \\ \mathbf{n}_2 \end{bmatrix}, \tag{6.42}$$

where $\mathbf{H}_1$ is

$$\mathbf{H}_1 = \begin{bmatrix} h_{11} & h_{21} \\ h_{21}^* & -h_{11}^* \end{bmatrix} \tag{6.43}$$

and $\mathbf{H}_2$ is

$$\mathbf{H}_2 = \begin{bmatrix} h_{12} & h_{22} \\ h_{22}^* & -h_{22}^* \end{bmatrix}. \tag{6.44}$$

Let us multiply

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \tag{6.45}$$

with

$$\begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_2 \end{bmatrix}^H \tag{6.46}$$

in order to get

$$\begin{aligned} \zeta_1 &= (|h_{11}|^2 + |h_{11}|^2 + |h_{11}|^2 + |h_{11}|^2)x_1 + n_\Sigma \\ \zeta_2 &= (|h_{11}|^2 + |h_{11}|^2 + |h_{11}|^2 + |h_{11}|^2)x_2 + n_\Sigma \end{aligned}, \tag{6.47}$$

where the resultant SNR is

$$\tau_{2\times 2} = \frac{(|h_{11}|^2 + |h_{11}|^2 + |h_{11}|^2 + |h_{11}|^2)E_b}{2N_o}. \tag{6.48}$$

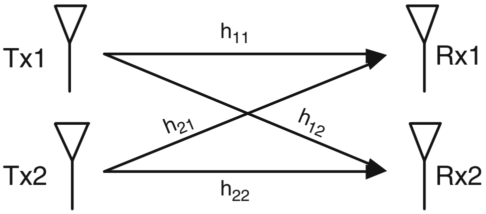


Fig. 6.10 MIMO 2×2 channel

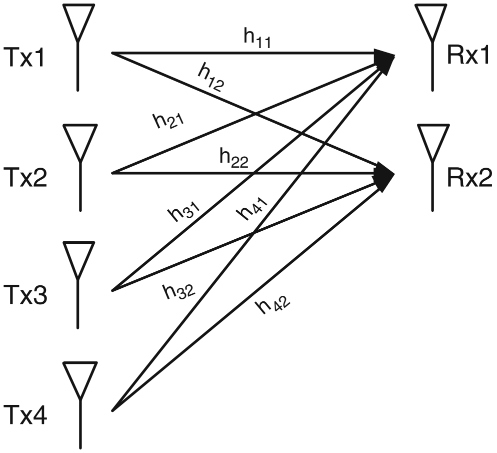


Fig. 6.11 MIMO 4 × 2 channel

As we see diversity gain is 4 but array gain is 2 due to the transmit power penalty: Resultant SNR is 3 dB less than 1 × 4 SIMO MRC system.

This 2 × 2 MIMO can be stacked to obtain 4 × 2 MIMO as seen in Fig. 6.11, in which data streams are doubled since four different signals are sent in each symbol and repeated in the second symbol. The received signals are now represented as below

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix}=\begin{bmatrix}\mathbf{H}_{11} & \mathbf{H}_{12}\\ \mathbf{H}_{21} & \mathbf{H}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}+\begin{bmatrix}\mathbf{n}_1\\ \mathbf{n}_2\end{bmatrix}, \tag{6.49}$$

where $\mathbf{H}_{ij}$ is a channel matrix as in the Alamouti scheme.

This scheme has a multiplexing gain of two and the resultant SNR suffers from interference due to the two simultaneous data streams. For instance, in the first branch SNR would be as follows:

$$\tau_{4\times 2}=\frac{(|h_{11}|^2+|h_{11}|^2+|h_{11}|^2+|h_{11}|^2)E_b}{I_3+I_4+4N_o}, \tag{6.50}$$

where I_3 and I_4 are interferences from x_3 and x_4. As a result, diversity gain can reach four if interference can be suppressed by maximum likelihood receiver, otherwise two with zero forcing decoder.

6.6.3 Spatial Multiplexing Gain

MIMO system is also used to increase data rate by spatial multiplexing. Spatial multiplexing creates independent signaling paths to send independent data and assumes accurate channel knowledge at the receiver.

The multiplexing gain decomposes MIMO channel into R parallel independent paths, which can increase data rate R-fold as compared with SISO case. R is called *multiplexing gain*.

Let us recall the MIMO model introduced in (6.3)–(6.5) at the beginning of this chapter:

$$\mathbf{y} = \mathbf{Hx} + \mathbf{n} \tag{6.51}$$

and singular value decomposition (SVD) of $\mathbf{H}$ is

$$\mathbf{H}_{M\times K} = \mathbf{U}_{M\times M} D_{M\times K} \mathbf{V}^{H}_{K\times K}, \tag{6.52}$$

where D is a diagonal matrix of singular values $\{\sigma_i\}$ of $\mathbf{H}$ and the number of non-zero eigenvalues are R, where $R \le \min(K,M)$.[2] As a result, the precoding at the transmitter with $\mathbf{V}$ and postcoding at the receiver with $\mathbf{U}^H$ as seen in Fig. 6.12 result in $\bar{\mathbf{y}}$,

$$\begin{aligned}
\bar{\mathbf{y}} &= \mathbf{U}^H \mathbf{y} \\
\bar{\mathbf{y}} &= \mathbf{U}^H (\mathbf{Hx} + \mathbf{n}) \\
\bar{\mathbf{y}} &= \mathbf{U}^H (\mathbf{U} D \mathbf{V}^H \mathbf{x} + \mathbf{n}) \\
\bar{\mathbf{y}} &= \mathbf{U}^H \mathbf{U} D \mathbf{V}^H \mathbf{V} \bar{\mathbf{x}} + \mathbf{U}^H \mathbf{n}) \\
\bar{\mathbf{y}} &= D\bar{\mathbf{x}} + \bar{n},
\end{aligned} \tag{6.53}$$

where $\bar{\mathbf{n}}$ and $\mathbf{n}$ are identically distributed. The result can be rewritten as follows

$$\bar{y}_i = \sigma_i \bar{x}_i + \bar{n}_i, \tag{6.54}$$

where $i \in \{1, \cdots, R\}$. If σ_i is small, there is high chance of error in that path. Notice that this optimal scheme assumes that $\mathbf{H}$ is known to the transmitter since independent data streams are modulated with $\mathbf{V}$. Previously, we have seen that the Jacobi rotation is typically used in real systems to diagonalize the channel correlation matrix with *estimated* $\mathbf{H}$. Without loss of generality, we continue our analysis with $\mathbf{H}$.

In the next sections, MIMO capacity is given with and without channel side information at the transmitter. This capacity has different interpretation for deterministic channels and fading channels since *deterministic MIMO channel* assumes that noise is nonrandom and additive white complex zero-mean Gaussian noise with covariance matrix $E(nn^H)$ equals $N_o\mathbf{I}_M$. On the other hand, *fading MIMO channel* assumes channel gain matrix $\mathbf{H}$ under flat fading, and so gains (h_{ij}) vary with time.

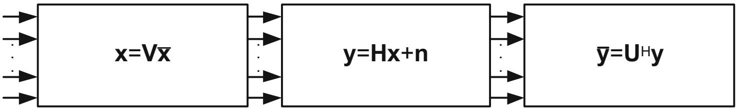


Fig. 6.12 MIMO – independent path

[2] Equality holds when there is rich scattering environment, otherwise it leads to high correlation among paths and R is small.

6.6.4 MIMO Capacity with CSI

As we recall from Chap. 2 for SISO, in deterministic MIMO channel, capacity is given in terms of mutual information

$$C = \max_{p(x)} I(\mathbf{x};\mathbf{y}), \tag{6.55}$$

where mutual information of output and input is $I(\mathbf{y};\mathbf{x}) = h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x})$ from which we can find $I(\mathbf{y};\mathbf{x}) = h(\mathbf{y}) - h(\mathbf{n})$ since $h(\mathbf{y}|\mathbf{x}) = h(\mathbf{n})$.

Remember from Chap. 2 that differential entropy of Gaussian random variable with variance σ^2 is $\frac{1}{2}\log_2(2\pi e\sigma^2)$ and differential entropy of multivariate Gaussian $\mathbf{S}$ is given by $\frac{1}{2}\log_2 \det[2\pi e R_{\mathbf{ss}}]$, where $R_{\mathbf{ss}}$ is covariance matrix.

Therefore, we can infer that noise has fixed entropy. To maximize the mutual information, we need to maximize the differential entropy of $\mathbf{y}$. Let us look at covariance matrix $E(\mathbf{y}\mathbf{y}^H)$ of $\mathbf{y}$,

$$E(\mathbf{y}\mathbf{y}^H) = \mathbf{H}R_{\mathbf{xx}}\mathbf{H}^H + N_o\mathbf{I}_M. \tag{6.56}$$

Mutual information of input and output is maximized when input and output is Gaussian. Thus, $I(\mathbf{x};\mathbf{y})$ is

$$I(\mathbf{x};\mathbf{y}) = \log_2[\mathbf{I}_M + \mathbf{H}R_{\mathbf{xx}}\mathbf{H}^H] \tag{6.57}$$

and capacity becomes

$$C = \log_2 \det[\mathbf{I}_M + \mathbf{H}R_{\mathbf{xx}}\mathbf{H}^H]\,\text{bps/Hz}, \tag{6.58}$$

As we can see, the capacity depends on how the power is distributed in the transmitter. Remember that total transmission power is bounded with P, where $\sum_1^K E(x_i^2) \leq P$. If $\mathbf{H}$ is known at the transmitter, an optimal allocation of transmit power is possible in order to maximize the capacity.

We can easily modify (6.58) to get

$$C = \max_{P_i} \sum_{i=1}^{R} \log_2\left(1 + \frac{\sigma_i^2 P_i}{\sigma^2}\right)\text{bps/Hz}, \tag{6.59}$$

where R is the rank of $\mathbf{H}$ and σ_i basically indicates the power allocation to an independent channel out of R independent channels. If $\tau_i = \sigma_i^2 P/\sigma^2$, C becomes

$$C = \max_{P_i} \sum_{i=1}^{R} \log_2\left(1 + \frac{\sigma_i^2 P_i \tau_i}{P}\right)\text{bps/Hz}, \tag{6.60}$$

where we can easily interpret that, if SNR is high, then the power allocated to the branch can be minimum P_i so that all branches get enough power linearly to maximize the capacity with the degrees of freedom in the channel, which is equivalent to the number of independent paths. If SNR is low then it is better to concentrate on one branch that has the largest SNR. As a result, there is a threshold τ_o value where P_i is selected according to $P/\tau_o - \sigma^2/\sigma_i^2$ and the resulting capacity then becomes

$$C = \sum_{i=1,\tau_i \geq \tau_o}^{R} \log_2\left(\frac{\tau_i}{\tau_o}\right) \text{bps/Hz}. \tag{6.61}$$

As it can be seen, it gives linear capacity increase with the increasing number of transmit antennas. Intuitively, it says that low-power transmission of data in many different channels is superior in capacity than high power transmission in one single channel.

Fading environment changes the channel gain (h_{ij}) with time, the capacity definition is rephrased as the average capacity over all realizations. If the channel is known at the transmitter, then all realizations are averaged over with optimal power allocation. The capacity, also known as ergodic capacity, $(\mathbf{E}[C])$ becomes

$$C = \mathbf{E}\left[\max_{P_i:\sum_i P_i \leq \bar{P}} \sum_i \log_2\left(1 + \frac{P_i \tau_i}{\bar{P}}\right)\right], \tag{6.62}$$

where $\tau_i = \sigma_i^2 \bar{\tau}$.

6.6.5 MIMO Capacity Without CSI

If the transmitter does not know the channel information, it cannot optimize its power for each branch. The best remaining strategy is to select equal transmit power. Hence, $R_{\mathbf{xx}}$ becomes $(P/K)\mathbf{I}_K$. This yields capacity to be

$$C = \log_2 \det\left[\mathbf{I}_M + \frac{\hat{\tau}}{K}\mathbf{H}\mathbf{H}^H\right] \text{bps/Hz} \tag{6.63}$$

and C becomes

$$C = \sum_{i=1}^{R} \log_2(1 + \tau_i/K) \, \text{bps/Hz}, \tag{6.64}$$

where $\tau_i = \sigma_i^2 \hat{\tau}$.

The equation converges to $C = \min(K,M)\log_2(1+\hat{\tau})$ if $\min(K,M)$ grows large. As a result, even in the absence of channel side information at the transmitter, capacity, grows linearly with the minimum number of transmitter and receiver.

We also see that if SNR is low, the number of transmit antennas does not have an impact on the capacity, since the power is concentrated on the transmit antennas that has the highest SNR.

In fading environment when channel side information is not available at the transmitter, the input covariance matrix is selected to maximize the capacity with transmission power constraint. Optimum case is when power is distributed equally among the transmit antennas as in the deterministic MIMO channel case. The ergodic capacity ($\mathbf{E}[C]$) becomes

$$C = \mathbf{E}[\log_2 \det[\mathbf{I}_M + \frac{\tau}{K}\mathbf{H}\mathbf{H}^H]]\text{bps/Hz}, \tag{6.65}$$

where the capacity also linearly scales with the $\min(K,M)$ as in the static case and note that ergodic capacity is a function of $E(\mathbf{H}\mathbf{H}^H)$. Remember that channel coding theorem requires a rate $\Re$, i.e. ergodic capacity does not indicate the error-free information rate. Another definition for capacity is *outage probability* for a given rate $\Re$ saying that

$$P_{\text{out}}(R) = P(C \leq \Re). \tag{6.66}$$

Notice that up to now we assume that channel side information is available at the *receiver*. If there is no channel side information at the receiver and transmitter, the capacity depicts a noteworthy characteristic with increasing SNR. Since unlike the previous cases in high SNR, capacity growth with SNR does not depend on degrees of freedom. On the other hand, in low and mid SNR, it grows linearly with degrees of freedom.

6.7 Space-Time Coding

Space-time coding introduces a coding scheme over time and space. Previously we investigated the coding schemes that only spans over time or frequency. If we follow the previously introduced MIMO system with K transmit and M receive antennas, input $\mathbf{X}$ is now a matrix that spans over T symbols and output $\mathbf{Y}$ becomes

$$\mathbf{Y}_{M\times T} = \mathbf{H}_{M\times K}\mathbf{X}_{K\times T} + \mathbf{N}_{M\times T}, \tag{6.67}$$

where $\mathbf{X}$ is encoded using one of the two techniques described below. Note also that channel is known only at the receiver.

6.7.1 Space-Time Block Coding (STBC)

STBC system seen in Fig. 6.13 operates on a block of input symbols to produce a matrix output in space and time. The encoder operates on a block of symbols to produce a $K \times T$ codeword ($\mathbf{X}$).

Figure 6.14 depicts the STBC codes. STBC codes are orthogonal[3] on the expense of reduced data rate. Orthogonality comes from the fact that transmission of an

[3] The Alamouti applies STBC. The rate is 1 since it takes two time-slots to transmit two symbols. This is a result of the perfect orthogonality between the symbols after receive processing. It is the only orthogonal STBC that achieves rate-1 without needing to sacrifice its data rate.

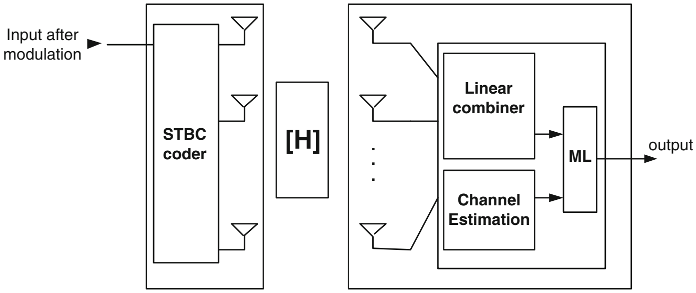


Fig. 6.13 STBC system: maximum-likelihood (ML) detection based only on linear processing at the receiver

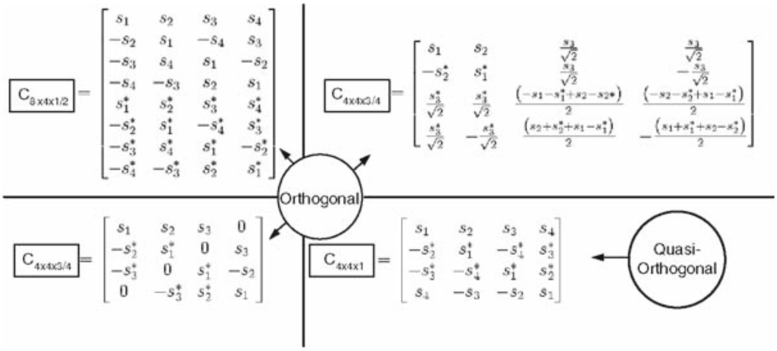

$C_{8x4x1/2} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ -s_2 & s_1 & -s_4 & s_3 \\ -s_3 & s_4 & s_1 & -s_2 \\ -s_4 & -s_3 & s_2 & s_1 \\ s_1^* & s_2^* & s_3^* & s_4^* \\ -s_2^* & s_1^* & -s_4^* & s_3^* \\ -s_3^* & s_4^* & s_1^* & -s_2^* \\ -s_4^* & -s_3^* & s_2^* & s_1^* \end{bmatrix}$
$C_{4x4x3/4} = \begin{bmatrix} s_1 & s_2 & \frac{s_3}{\sqrt{2}} & \frac{s_3}{\sqrt{2}} \\ -s_2^* & s_1^* & \frac{s_3}{\sqrt{2}} & -\frac{s_3}{\sqrt{2}} \\ \frac{s_3^*}{\sqrt{2}} & \frac{s_3^*}{\sqrt{2}} & \frac{(-s_1-s_1^*+s_2-s_2*)}{2} & \frac{(-s_2-s_2^*+s_1-s_1^*)}{2} \\ \frac{s_3^*}{\sqrt{2}} & -\frac{s_3^*}{\sqrt{2}} & \frac{(s_2+s_2^*+s_1-s_1^*)}{2} & -\frac{(s_1+s_1^*+s_2-s_2^*)}{2} \end{bmatrix}$
Orthogonal
$C_{4x4x3/4} = \begin{bmatrix} s_1 & s_2 & s_3 & 0 \\ -s_2^* & s_1^* & 0 & s_3 \\ -s_3^* & 0 & s_1^* & -s_2 \\ 0 & -s_3^* & s_2^* & s_1 \end{bmatrix}$
$C_{4x4x1} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ -s_2^* & s_1^* & -s_4^* & s_3^* \\ -s_3^* & -s_4^* & s_1^* & s_2^* \\ s_4 & -s_3 & -s_2 & s_1 \end{bmatrix}$
Quasi-Orthogonal

Fig. 6.14 STBC code design: $C_{T\times K\times R}$ indicates the T symbols, K transmit antennas, and R is rate. (Source: en.wikipedia.org)

antenna over T symbol period is orthogonal to transmission of another antenna over T symbol period. This way complexity in the receiver is lower since decoder is linear and optimal. Quasi-orthogonal STBC is possible with higher data rate but with increased ISI.

In the decoding side, receiver assumes channel side information. Received vector is multiplied with $\mathbf{H}^H$ to achieve a modified received signal vector. Since the codeword matrix is orthogonal, the first term in the linear combining equals the identity matrix $\mathbf{I}$ multiplied by a factor of the fading gains. Then maximum-likelihood estimator can easily decode it with a complexity of $O(\min\{K,M\})$.

STBC has some limits. Unlike the Alamouti scheme that achieves full diversity, a general STBC for any arbitrary number of transmit antennas exists only for code rate of 1/2 and only two additional codes are known, with a higher code rate than 1/2 for more than two antennas. For instance, with three-antenna STBC, maximum rate can be achieved with a code rate 1/2 and 3/4. The highest rate is given by

$$\text{Rate} = \frac{K+3}{(2K)||(2K+2)}. \tag{6.68}$$

If X is codeword and $\hat{X}$ is received signal, maximum possible diversity order of $K \times M$ is achieved when $\mathbf{B}(X,\hat{X}) = X - \hat{X}$ is full rank. $\mathbf{B}(X,\hat{X})$ is given by

$$\mathbf{B}(X,\hat{X}) = X - \hat{X} = \begin{bmatrix} x_1^1 - \hat{x}_1^1 & x_2^1 - \hat{x}_2^1 & \cdots & x_T^1 - \hat{x}_T^1 \\ x_1^2 - \hat{x}_1^2 & x_2^2 - \hat{x}_2^2 & \cdots & x_T^2 - \hat{x}_T^2 \\ \vdots & \vdots & \cdots & \vdots \\ x_1^K - \hat{x}_1^K & x_2^K - \hat{x}_2^K & \cdots & x_T^K - \hat{x}_T^K \end{bmatrix} \tag{6.69}$$

and if rank is R then the diversity order is RM.

STBC offers only diversity gain (compared with single-antenna schemes) but not coding gain, since the redundancy purely provides diversity in space and time. This limitation may be alleviated by concatenation. STBC together with Trellis coding introduce diversity gain as well as coding gain.

6.7.2 Space-Time Trellis Coding (STTC)

STTC provides coding gain similar to Trellis coding for SISO in contrast to STBC. STTC also provides diversity gain similar to STBC, but with high complexity in decoding with $O(m^{\min\{K,M\}})$, where m is modulation (e.g., $m = 4$ for QPSK). It relies on Viterbi decoder at the receiver. Coding performance is quantified by diversity advantage and coding advantage.

Figure 6.15 shows a STTC system for K transmit antennas, where input $(c_0, c_1, \ldots, c_t, \ldots)$ is sent to m shift registers. Each input experiences a generator sequence and output of the encoder becomes the following

$$\left(x_t^i = \sum_{m}^{k=1} \sum_{\nu_k}^{j=0} g_{j,i}^k c_{t-j}^k \right) \bmod m, \tag{6.70}$$

where generator sequences are given by

$$g^k = \begin{bmatrix} (g_{0,1}^k, g_{0,2}^k, \cdots, g_{0,K}^k), \\ (g_{1,1}^k, g_{1,1}^k, \cdots, g_{1,K}^k), \\ \cdots \\ (g_{\nu_k,1}^k, g_{\nu_k,2}^k, \cdots, g_{\nu_k,K}^k) \end{bmatrix} \tag{6.71}$$

for $k \in \{1, \ldots, m\}$. Total memory required for the encoder is $\upsilon = \sum_{k=1}^{m} \nu_k$, since 2^{υ} states are required.

STTC operates on one input symbol to produce vector outputs. Decoding is performed via ML sequence estimation. The redundant copies of a Trellis code are mapped in time and space.

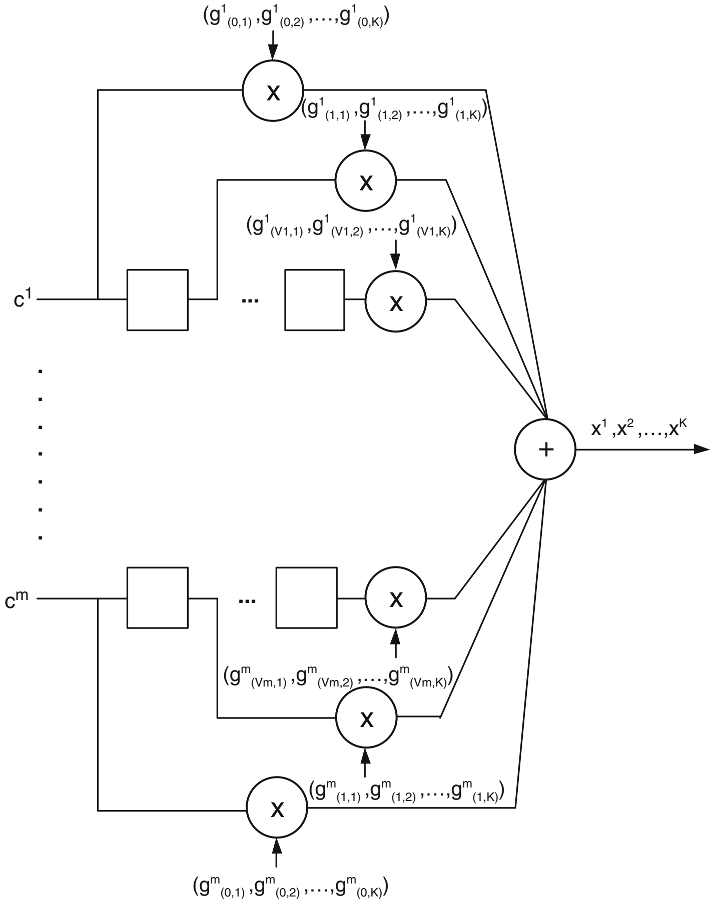

(g1(0,1), g1(0,2), …, g1(0,K))
x
(g1(1,1), g1(1,2), …, g1(1,K))
x
(g1(V1,1), g1(V1,2), …, g1(V1,K))
c1
…
x
+
x1, x2, …, xK
cm
…
x
(gm(Vm,1), gm(Vm,2), …, gm(Vm,K))
x
(gm(1,1), gm(1,2), …, gm(1,K))
x
(gm(0,1), gm(0,2), …, gm(0,K))

Fig. 6.15 STTC system

To design optimum STTC, the number of transmit antennas and the number of states are taken into account. Best code minimizes the error probability of $\mathbf{B}(X,\hat{X})$ as given in (6.69).

Code design depends on the rank (R) of $\mathbf{B}$ and the number of received antennas M. If $RM < 4$, the rank and determinant criteria is selected, otherwise the trace criteria is preferred. Determinant criteria maximizes the minimum product, $\Pi\alpha_i$ of $\mathbf{BB}^H$ matrix along the pairs of distinct codewords with the minimum rank. Trace criteria maximizes the minimum trace $\sum\alpha_i$ of $\mathbf{BB}^H$ matrix among all pairs of distinct codewords. Achievable rank R is given in Table 6.4 as a function of υ and K transmit antennas.

Figure 6.16 compares the STTC with a simple diversity technique. Simple diversity is chosen as delay diversity, in which each branch in the transmitter transmits with a delay as seen in Fig. 6.17. Notice that STTC, which is more complex than

Table 6.4 Achievable rank

	$K=2$	$K=3$	$K=4$	$K=5$	$K\geq 6$
$\upsilon=2$	2	2	2	2	2
$\upsilon=3$	2	2	2	2	2
$\upsilon=4$	2	3	3	3	3
$\upsilon=5$	2	3	3	3	3
$\upsilon=6$	2	3	4	4	4

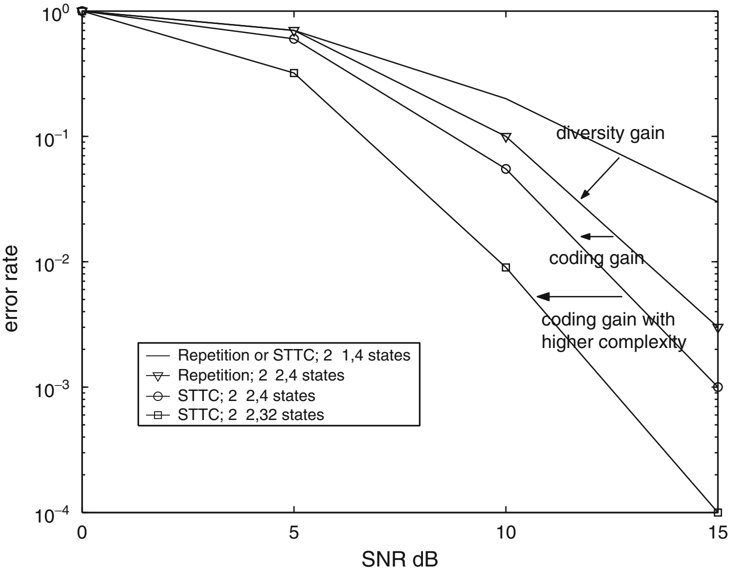


Fig. 6.16 Comparison of STTC with delay diversity and simple coding (repetition coding): Notice that diversity gain changes the slope of the error curve, and the coding gain results in a horizontal shift

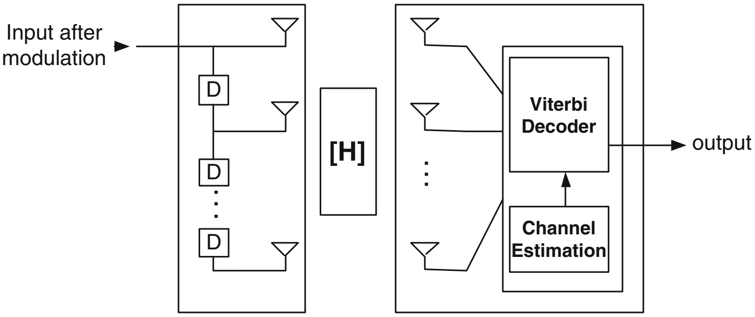


Fig. 6.17 Delay diversity system: Consider the same delay in all paths that transfer the flat fading channel into a channel with intersymbol interference (ISI). Since the fading gains in $[\mathbf{H}]$ are random, the overall channel is a random intersymbol interference channel. Maximum likelihood sequence estimator (e.g., Viterbi) can provide optimal decoding with full diversity gain

delay diversity does not increase the performance with only one receive antenna. Coding gain on the other hand affects the horizontal shift, and the performance can be further improved with more complexity in the decoding.

6.8 MIMO BLAST Transceiver

MIMO BLAST[4] transceiver provides spatial-multiplexing transmission with multiple concurrent data streams. The two most important techniques that decouple the data streams are diagonal BLAST (D-BLAST) introduced by Foschini and vertical BLAST (V-BLAST) introduced by Foschini et al.

D-BLAST provides a diagonally layered coding architecture as presented in Fig. 6.18. Compare it with typical serial encoding MIMO system in Fig. 6.19. D-BLAST disperses the code blocks across diagonals in space and time. This achieves theoretical rates (approaching 90% Shannon capacity) presented above in the rich scattering environment.

D-BLAST introduces redundancy between substreams with the use of intersubstream block coding, which are organized along diagonals in space-time. D-BLAST has layered architecture; each layer is decoded diagonal wise as seen in Fig. 6.20. During decoding, not detected layers are nulled with channel information and previously detected layers are canceled by subtracting. Notice that the decoding is iterative and inefficient because of not used slots.

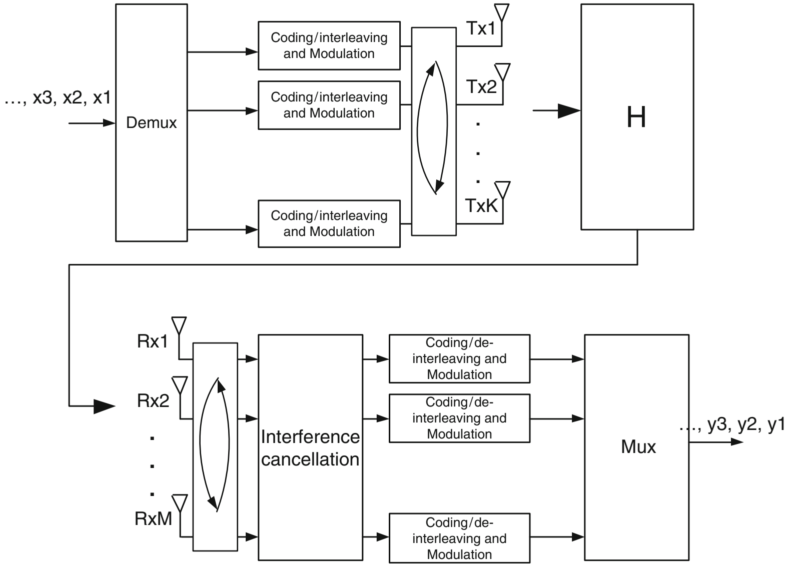

..., x3, x2, x1
Demux
Coding/interleaving and Modulation
Coding/interleaving and Modulation
Coding/interleaving and Modulation
Tx1
Tx2
TxK
H
Rx1
Rx2
RxM
Interference cancellation
Coding/de-interleaving and Modulation
Coding/de-interleaving and Modulation
Coding/de-interleaving and Modulation
Mux
..., y3, y2, y1

Fig. 6.18 MIMO D-BLAST transceiver

[4] BLAST is an acronym for Bell Labs layered Space-Time.

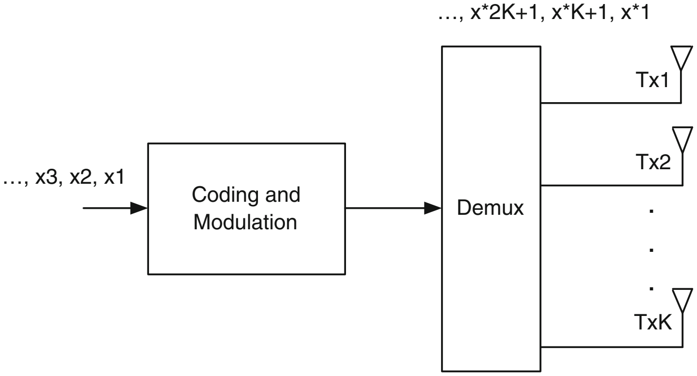


Fig. 6.19 Serial encoding

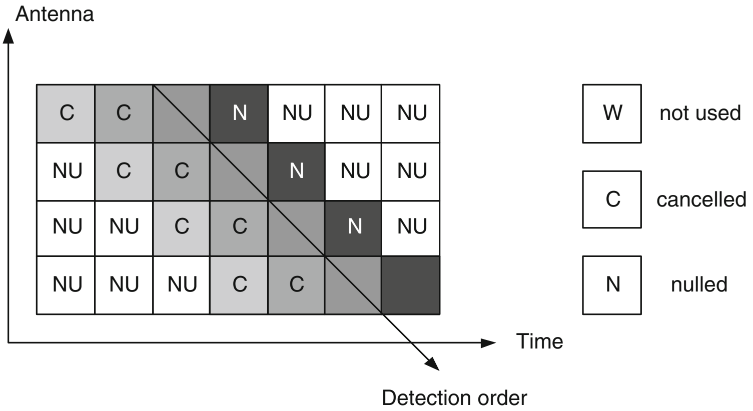
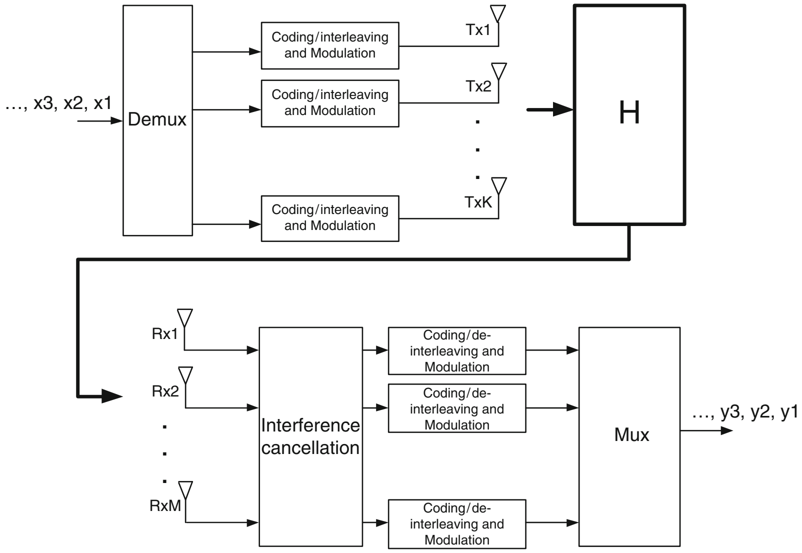


Fig. 6.20 D-BLAST decoding

V-BLAST on the other hand uses vector encoding process, which is a pretty much simplified and efficient version of D-BLAST. Figure 6.21 shows the high-level block diagram of a V-BLAST system. A data stream is demultiplexed into K multiple streams, where each transmitter operates at symbol rate $1/T$ symbols/s with synchronized symbol timing. The power is proportional to $1/K$ so that the total power is constant and independent of the number of transmitting antennas.

There is no intersubstream coding as in D-BLAST. Vectorized transmission is shown in Fig. 6.22. The main idea behind V-BLAST detection is symbol cancelation and linear nulling with zero-forcing or MMSE. Conceptually each substream is considered as the desired signal and the remainder are considered as interferers. Nulling is done with giving weight to received signal to satisfy MMSE or ZF criteria. If received signal is $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$, zero nulling vector is

$$\mathbf{g}_i^{\mathrm{T}}\mathbf{H}_j = \delta_{i,j}, \tag{6.72}$$



…, x3, x2, x1
Demux
Coding/interleaving and Modulation
Coding/interleaving and Modulation
Coding/interleaving and Modulation
Tx1
Tx2
TxK
H
Rx1
Rx2
RxM
Interference cancellation
Coding/de-interleaving and Modulation
Coding/de-interleaving and Modulation
Coding/de-interleaving and Modulation
Mux
…, y3, y2, y1

Fig. 6.21 MIMO V-BLAST transceiver

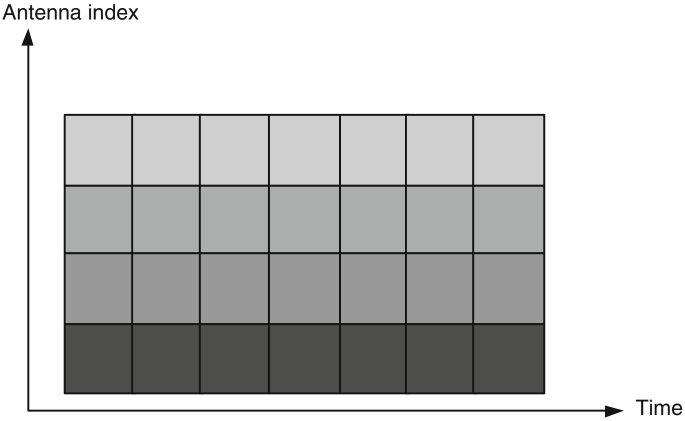


Fig. 6.22 V-BLAST decoding

where δ is the Kronecker delta and $\mathbf{H}_j$ is the jth column of $\mathbf{H}$. The decision is given by

$$y_i = \mathbf{g}_i^{\mathrm{T}} \mathbf{y}, \tag{6.73}$$

where $\hat{x}_i$ is the slice of y_i.

Better than linear nulling, iterative approach gives superior performance in which interference from already-detected components of $\mathbf{x}$ is subtracted out from the

Table 6.5 Comparison of STBC, STTC, V-BLAST, and D-BLAST

	Advantage	Disadvantage
STBC	Simple, diversity gain	No coding gain
STTC	Diversity and coding gain	Complex
V-BLAST	Simple spatial multiplexing, no need coding, high throughput	No diversity gain, $M \geq K$
D-BLAST	No requirement for $M \geq K$, close to theoretical capacity	Complex coding

Note that there is no extra power and bandwidth requirement in all the schemes

received signal vector. Ordering of iterative process is important, where the strongest symbol is detected first. The resultant SNR becomes

$$\tau_i = \frac{< |x_i|^2 >}{\sigma^2 ||\mathbf{g}_i||^2}, \tag{6.74}$$

for $i \in \{1, \cdots, K\}$. V-BLAST has demonstrated spectral efficiencies of 20–40 bps/Hz with the prototype when introduced first time in Bell Labs.

Table 6.5 shows the comparison of diversity and spatial-multiplexing schemes of MIMO. They can exist in the same system with careful balancing. Optimal diversity gain showed by Zheng and Tse indicates that any coding scheme of block length larger than $K+M-1$ with multiplexing gain m is precisely $(K-m)(M-m)$ for i.i.d Rayleigh slowly-fading channel. This indicates that $m \times m$ is used for multiplexing gain and $(K-m) \times (M-m)$ is used for diversity gain.

6.9 MIMO with HARQ

We know that HARQ enhances performance by combining the previous packet with the retransmitted packet. Similar to SIMO, MRC in SISO with HARQ also achieves the maximum SNR. However, when MIMO system with HARQ is used, how to combine multiple received vector is not clear because of the interference coming from different antennas.

Previously, utilizing ZF and MMSE decoding is proposed to convert a MIMO system into a SISO so that MRC can be performed with retransmitted copies. There are other combining proposals depicted in Table 6.6 to address the key design factors. Let us first introduce the MIMO system with HARQ as follows

$$\mathbf{y}_i = \mathbf{H}_i \mathbf{x}_i + \mathbf{n}_i \tag{6.75}$$

for $i \in \{1, \ldots, N\}$, where N is the number of retransmissions.

Distance-level combining (DLC) compares the distance $||\mathbf{y}_i - \mathbf{H}_i \hat{\mathbf{x}}_i||$ for each i and calculates two distances for $\hat{\mathbf{x}}_i^j$ for $j \in \{0, 1\}$. The minimum is taken for each j

Table 6.6 Comparison of combining schemes for MIMO with HARQ

	DLC	MRC	CASLC	BLC	MRC-equalized
Performance	Optimal	Optimal	Optimal	Suboptimal	Suboptimal
Scalability	Yes	Yes	Yes	Yes	Yes
Memory Size	High	Small	Small	Small	Medium
Complexity	High	Medium	Low	Medium	Low
HARQ-CC	O	O	O	O	O
HARQ-IR	X	X	X	O	X

Key design concepts in combining are as follows: decoding performance, which is a metric to measure how full the relevant information is used without any loss in LLR; scalability, which is reusability of basic decoder regardless of the number of retransmission; memory size, which is receiver memory size reserved for retransmission

and the difference gives the log-likelihood ratio (LLR)[5] of ML decoding as soft-bit information metric. On the other hand, maximum ratio combining method first performs MRC operation in the receiver and then calculates the LLR. However, bit-level combining (BLC) first calculates the LLR from individual received signals and then sums up those to get the final combined LLR. Recently proposed concatenation-assisted symbol-level combining (CASLC) performs *QR* decomposition on the channel **H** and calculates the LLR using $||Qy - Rx||^2$. Maximum ratio combining of equalized symbols first performs linear equalization, e.g., ZF, then employs MRC to calculate the LLR. Note that BLC is optimal for HARQ-IR and CASLC shows better performance for HARQ-CC.

6.10 Multiuser MIMO – SDMA

Multiuser MIMO[6] (MU-MIMO) is a term to define a system in which a MIMO transmission aims multiple users concomitantly. MU-MIMO receivers are more complex than single-user MIMO (SU-MIMO) because of simultaneous detection of the signals from all users.

In cellular-type deployment, a base station reaches subscribers via downlink MIMO broadcast channel, and mobile subscribers reach the base station via uplink MU-MIMO channel as seen in Fig. 6.23. This ordering in downlink and uplink is managed by a multiple access scheme, termed *SDMA* (space-division multiple access).

[5]

$$\text{LLR} = \ln \frac{\Pr\{b_z = 1|\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{H}_1, \ldots, \mathbf{H}_N\}}{\Pr\{b_z = 0|\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{H}_1, \ldots, \mathbf{H}_N\}}, \tag{6.76}$$

where b_z is transmit bit sequence $b \in \{0,1\}^{mK}$ in M-QAM modulation with $M = 2^m$.

[6] “Similar to the relationship between OFDM and OFDMA, MU-MIMO is the extended technique of MIMO to be used as a multiple access method...”.

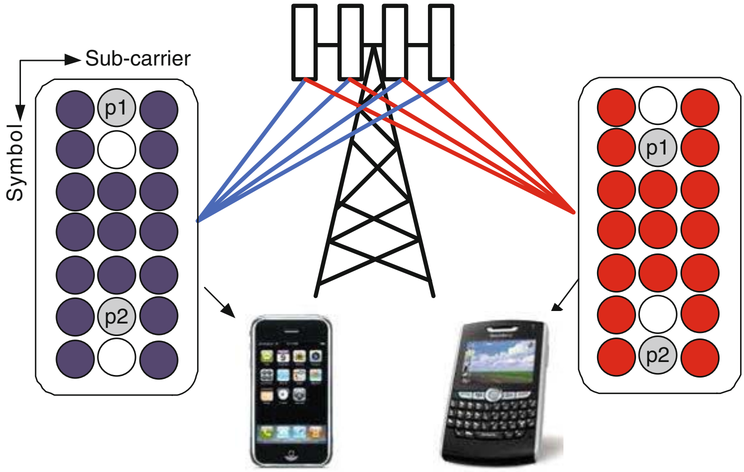


Fig. 6.23 Multiuser MIMO: There is no 3dB power penalty

The capacity benefits of MU-MIMO can be even greater than SU-MIMO. Previously, we see that linear capacity gain is increased with $\min(T_B, T_M)$, where T_B denotes antennas in the base station and T_M denotes number of antennas in the mobile subscriber. In multiuser MIMO, $T_B >> T_M$ since there are small number of antennas at each mobile. But, now the capacity increases linearly with $\min(T_B, m.T_M)$, where m is the number of mobiles. Thus, having a large number of mobiles can reduce the number of antennas at each mobile. This is one of the key consideration for space-limited mobile devices.

For uplink and downlink MU-MIMO, channel state information is required at the transmitter and receivers. In uplink, multiuser detection or successive interference cancelation can be used to achieve the capacity. The coding capacity is similar to SU-MIMO. Downlink MU-MIMO is an open problem, and recently there has been progress with so called "dirty paper coding" to achieve the capacity of the MIMO with fixed channel. MU-MIMO is being considered for WiMAX, LTE, and other next generation technologies.

6.11 Cooperative MIMO and Macrodiversity

So far our assumption was to consider antennas co-located in the same device but separated with a distance around half the wavelength. Now, consider the case where transmit antennas are not co-located and separated with a long distance as in Fig. 6.24. This type of situation may be leveraged to implement a cooperative MIMO system where information exchange can be created among nodes to create a virtual antenna array. In other words, nodes close together on the transmit side can create a multiple-antenna transmitter with information exchange, and nodes near each other in the receiver side can do this to create multiple-antenna receiver.

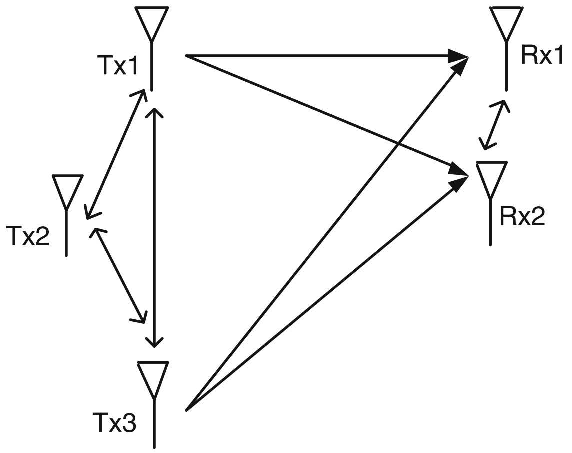


Fig. 6.24 Macrodiversity with two transmit antennas separated with long distance. Possible transmit antennas are in different cell sites. More than one cell cite can transmit and a receiver may have more than one co-located receive antennas

Cooperative MIMO has performance advantages in diversity, multiplexing, and beamforming, since now each node has an uncorrelated channel to each receiver.

Macrodiversity is a simple version of cooperative MIMO where there is no distributed antenna structure in the mobile station. Base stations participate to form a virtual antenna array for downlink in order to transmit information to the mobile stations. Also, mobile station's uplink transmission is received by multiple base stations. Hence, received uplink signals are aggregated in one location to perform the decoding. Of course, this requires extensive synchronization at the network to maintain the transmission ordering as well as CSI feedback. Macrodiversity, introduced in IEEE 802.16e, increases the signal quality and offers a seamless handover structure.

6.12 Other Smart Antenna Techniques

The use of multiple antennas is also considered in the context of beamforming, which are suitable in low-scattering environment but indoors or in urban deployments. In beamforming, an array of antenna is used to create a directional beam pattern.

Directional beamforming (aka direction of arrival beamforming) creates the radiation pattern of the antenna array by adding constructively the phases in the desired direction and destructively in the other directions to null the interference. Unlike MIMO beamforming explained above, directional beamforming creates a beam toward a physical direction. Channel side information at the transmitter is the angle of arrival and delay in each branch. MUSIC (Multiple Signal

Classification), ESPRIT (estimation of signal parameters via rotational invariant techniques) algorithms, Matrix Pencil method, or their derivatives are well-known beamforming techniques. They involve finding a spatial spectrum of the antenna/sensor array, and calculate the direction of arrival (DOA) from the peaks of this spectrum.

Switched fixed beam array is another example with several fixed beams that are selected accordingly. The system selects one or more at a given time. At the same time, it also nulls the interfering signals. More advanced beamformers are available: smart antennas with adaptive beam steering is one example that control the gain/phase with digital signal processing.

6.13 Application: IEEE 802.11n

The desire for higher rate in WLAN is being addressed in IEEE 802.11n standard.[7] The main objective of IEEE 802.11n is to achieve higher rates with mandatory interoperability requirement to legacy 802.11a/g systems. In IEEE 802.11n, MIMO-OFDM concept is introduced with several other features such as STBC, LDPC, adaptive beamforming, frame aggregation, block acknowledgement, and MAC header suppression.

Three modes of operation is introduced: legacy mode, mixed mode, and green field mode. In legacy mode, 802.11n uses only one antenna and the same burst structure of 802.11a/g in order to communicate to 802.11a/g systems. In mixed mode, 802.11n nodes communicate via MIMO-OFDM among each other, but if a receiver is legacy node then 802.11n can perform combining as in SIMO. To support MIMO operation and also preserve backward compatibility, additional preamble in PLCP header after legacy preamble for channel estimation and synchronization may be introduced.

A MIMO-OFDM transmitter with K transmit antennas is shown in Fig. 6.25. The randomizer is utilized to scramble incoming bits in order to avoid the occurrence of long zeros and ones. The scrambled bits are sent to n FEC encoders, where n equals

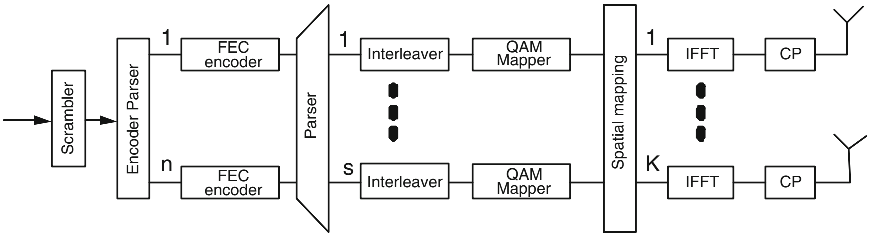


Fig. 6.25 MIMO-OFDM transmitter in IEEE 802.11n

[7] The IEEE in January 2006 confirmed the draft specification for the next-generation 802.11n WiFi standard.

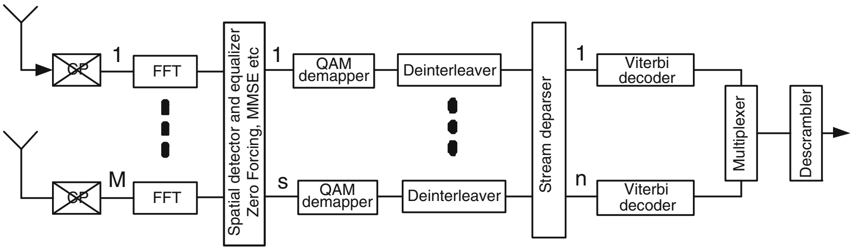


Fig. 6.26 MIMO-OFDM receiver in IEEE 802.11n for hard decoding. In case of soft decoding, first deinterleaving and decoding is performed

one for 1 × 1 and 2 × 2 and two for 3 × 3 and 4 × 4 systems. FEC encoder starts in zero state after encoding and achieves a basic coding rate of $\frac{1}{2}$. Other coding rates are achieved with puncturing. The output is mapped to s spatial stream with stream parser. Bits in each stream parser is interleaved and mapped to constellation point. After that, spatial mapper distributes the complex symbols to the K transmit chains. For a 20-MHz operation, 52 subcarriers are used for data out of 64 and the remaining subcarriers are used for pilot and guard interval. N point IFFT is taken and cylic prefix of $N/4$ in length is added before upconversion to radio frequency for transmission via K antennas.

A MIMO-OFDM receiver with M receive antennas is shown in Fig. 6.26. The signals in each receiver branch is downconverted and sampled with 50 ns, which is the maximum sampling duration. A standard OFDM receiver is used in each chain and a spatial demapper collects deinterleaved signals from s paths and multiplexes them into n Viterbi decoders. After decoding, descrambling is performed to rearrange bits.

In 802.11n, as in 802.11b/g, different rates can be achieved with various modulation and coding schemes as follows;[8]

$$\begin{aligned} \mathrm{Data-Rate} = {} & \mathrm{channel} \\ & \times \frac{\mathrm{data-subcarrier}}{\mathrm{total-subcarrier}} \\ & \times \mathrm{constellation-size} \\ & \times \mathrm{coding-rate} \\ & \times 1 - \frac{\mathrm{guard-time}}{\mathrm{symbol-time}} \\ & \times \mathrm{spatial-streams} \end{aligned} . \tag{6.77}$$

MIMO can increase the data rate multifold based on the number of spatial streams. Table 6.7 depicts the data rates for IEEE 802.11n, where the number of spatial streams is taken as the $\min(K,M)$. Other techniques such as beamforming, LDPC, and STBC are possible to increase the SNR but the rate calculation becomes nonlinear.

[8] For example, maximum PHY data rate (600 Mbps) in 802.11n is achieved with 5/6 coding rate, given that the system operates in 40-MHz channel, 108 subcarriers out of 128 are used for data transmission with 64QAM modulation, and the guard time and symbol duration are 400 ns and 4 μs respectively.

Table 6.7 MIMO-OFDM data rate table for IEEE 802.11n, where the number of subcarriers is 64 for 20 MHz and 128 for 40 MHz

Data rate (Mbps) with 800 ns CP	Data rate (Mbps) with 400 ns CP	Modulation	coding rate	Bandwidth (MHz)	Number of data subcarriers	s
6.5	7.2	BPSK	1/2	20	52	1
65	72.2	64QAM	5/6	20	52	1
13	14.4	BPSK	1/2	20	52	2
130	144	64QAM	5/6	20	52	2
19.5	21.7	BPSK	1/2	20	52	3
195	216.7	64QAM	5/6	20	52	3
26	28.9	BPSK	1/2	20	52	4
260	288.9	64QAM	5/6	20	52	4
13.5	15	BPSK	1/2	40	108	1
135	150	64QAM	5/6	40	108	1
27	30	BPSK	1/2	40	108	2
270	300	64QAM	5/6	40	108	2
40.5	45	BPSK	1/2	40	108	3
405	450	64QAM	5/6	40	108	3
54	60	BPSK	1/2	40	108	4
540	600	64QAM	5/6	40	108	4

For example, for 65 Mbps raw rate, the higher layer effective throughput is about 50 Mbps and the MAC overhead is 25%

Also, fallback mode of IEEE 802.11n to lower rates is improved with the introduction of Fast Modulation and Coding Scheme (MCS) feedback mechanism. Fast MCS installs explicit per-packet feedback to recommend the transmission speed for the next packet so that tracking rapid changes in the channel is addressed.

6.14 Summary

MIMO technology is being used in all next generation communication systems, especially in WiMAX and 4G systems. We described the basics of MIMO, including space-time coding, spatial multiplexing, and beamforming. Highlights are noted below:

- MIMO may either achieve boosted signal strength or higher data rates, where these two modes can be used in alternate order within the same cell. Theoretical capacity limit introduced by Shannon can linearly be increased with the number of transmit and receive antenna pairs. MIMO works best in rich-scattering environment and requires minimum RF coupling between spatially separated antennas.
- MIMO space-time coding aims to combat multipath fading with diversity gain. Same signal is coded differently and transmitted from each antenna in order to leverage the independent paths observed because of the rich fading. Then, in the receiver they are combined to get a better signal.

- MIMO spatial multiplexing aims to increase the spectral efficiency by sending independent data streams so that data rate is increased. Also, channel information helps to adjust the transmitter information with respect to changing channel conditions. Certain codebooks are defined for each standard in which channel information is sent back to transmitter with a signal that specifies the entry. Also in TDD systems, channel reciprocity is leveraged to obtain the channel information.
- MIMO transmission can be two types: multiuser and single-user. Multiuser MIMO carries information of different users in one transmission and single user MIMO carries information of only one user in a given transmission.

Interested readers can refer to Bibliography for more details.

References

1. Tse, D., Viswanath, P., *Fundamentals of Wireless Communication,* Cambridge University Press, Cambridge, 2005.
2. Goldsmith, A., *Wireless Communications,* Cambridge University Press, Cambridge, 2005.
3. Paulraj, A., Nabar, R., Gore, D., *Introduction to Space-Time Wireless Communications,* Cambridge University Press, Cambridge, 2003.
4. Biglieri, E., *MIMO Wireless Communications,* Cambridge University Press, Cambridge, 2007.
5. Agrawal, D., Naguib, A., Seshadri, N., Tarokh, V., "Space-time coded OFDM for high data-rate wireless communication wideband channels," *IEEE Conference Proceedings VTC,* pp. 2232–2236, 1998.
6. Foschin, G. J., Gans, M. J., "On limits of wireless communication in a fading environment when using multiple antennas," *Wireless Personal Communications,* vol. 6, pp. 311–335, 1998.
7. Gesbert, D., Naguib, A., Shafi, M., Shiu, D. S., Smith, P., "Theory to practice: an overview of MIMO space-time coded wireless systems," *IEEE Journal on Selected Areas in Communication,* vol. 21, pp. 281–302, 2003.
8. Foschini, G., Golden, G., Valenzuela, R., Wolniasky, P., "Detection algorithm and initial laboratory results using the V-BLAST space-time communication architecture," *Electronics Letters,* vol. 35, pp. 14–15, 1999.
9. Farrokhi, F. R., Lozano, A., Valenzuela, R., "Lifting the limits on high-speed wireless data access using antenna arrays," *IEEE Communications Magazine,* pp. 156–162, 2001.
10. Proakis, J. G., *Digital Communications,* 4th edition, McGraw Hill, New York, 2000.
11. Stuber, G. L., *Principles of Mobile Communication,* 2nd edition, Kluwer, Boston, 2001.
12. Foschini, G., Golden, G., Wolniansky, P., Valenzuela, R., "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," *International Symposium on Signals, Systems, and Electronics,* pp. 295–300, 1998.
13. Bisla, B., Eline, R., Franca-Neto, L. M., "RF system and circuit challenges for WiMAX," *Intel Technology Journal,* vol. 8, no. 3, pp. 189–200, 2004.
14. Viswanathan, S., "Tutorial on 802.11n PHY layer:Part 3," *Wireless Net Design Line.* `http://www.wirelessnetdesignline.com/howto/wlan/199703122.`
15. Ergen, M., Varaiya, P., "Formulation of distributed coordination function of IEEE 802.11 for asynchronous networks: mixed data rate and packet size," *IEEE Transactions on Vehicular Technology,* vol. 57, no. 1, pp. 436–447, 2008.
16. Pan, J.-L., Olesen, R., Grieco, D., Yen, S., "Efficient feedback design for MIMO SC-FDMA systems," *IEEE VTC,* Spring 2007.

17. Teletar, E., “Capacity of multi-antenna Gaussian channels,” *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.
18. Lebrun, G., Gao, J., Faulkner, M., “MIMO transmission over a time-varying channel using SVD,” *IEEE Transactions on Wireless Communication*, vol. 4, no. 2, pp. 757–764, 2005.
19. Hen. I., “MIMO architecture for wireless communication,” *Intel Technology Journal*, vol. 10, no. 2, pp. 157–165, 2006.
20. Jang, E. W., Lee, J., Lou, H.-L., Cioffi, J. M., “Optimal combining schemes for MIMO systems with hybrid ARQ,” *IEEE International Symposium on Information Theory*, pp. 2286–2290, June 2007.
21. Jang, E. W., Lee, J., Song, L., Cioffi, J. M., “Concatenation-assisted symbol-level combining scheme for MIMO systems with HARQ,” *IEEE Global Telecommunications Conference*, pp. 3275–3279, Nov. 2007.
22. Kim, J., Heath, R. W., Powers, E., “Receiver designs for alamouti coded OFDM systems in fast fading channels,” *IEEE Transactions Wireless Communications*, vol. 4, no. 2, pp. 550–559, 2005.
23. Barry, J., Lee, E., Messerschmitt, D. G., *Digital Communication*, Third Edition, Kluwer, Dordrecht, 2003.
24. Wolniansky, P., Foschini, G., Golden, G., Valenzuela, R., “V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel,” *Signals, Sytems, and Electronics*, 1998.
25. Sendonaris, A., Erkip, E., Aazhang, B., “User cooperation diversity – Part I: System description,” *IEEE Transactions on Communications*, vol. 51, no. 11, pp. 1927–1938, 2003.
26. Seshadri, N., Winters, J. H., “Two signalling schemes for improving the error performance of frequency-division-duplex (FDD) transmission systems using transmitter antenna diversity,” *International Journal of Wireless Information Networks*, vol. 1, pp. 49–60, 1994.
27. Stefanov, A., Erkip, E., “Cooperative coding for wireless networks,” *Proceedings of International Workshop on Mobile and Wireless Communication Networks*, pp. 273–277, September 2002.
28. Stridh, R., Ottersten, B., Karlsson, P., “MIMO channel capacity of a measured indoor radio channel at 5.8 GHz,” *Proceedings of Asilomar Conference on Signals, Systems, and Computers*, vol. 1, pp. 733–737, November 2000.
29. Tarokh, V., Jafarkhani, H., Calderbank, A. R., “Space-time block codes from orthogonal designs,” *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1456–1467, 1999.
30. Vishwanath, S., Jindal, N., Goldsmith, A., “Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, 2003.
31. Zheng, L., Tse, D. N. C., “Diversity and multiplexing: a fundamental trade-off in multiple-antenna channels,” *IEEE Transactions on Information Theory*, vol. 49, no. 5, pp. 1073–1096, 2003.
32. Touzni, A., Fijalkow, I., Larimore, M. G., Treichler, J. R., “A globally convergent approach for blind MIMO adaptive deconvolution,” *Signal Process.*, vol. 49, no. 6, pp. 1166–1178.
33. Foschini, G. J., Gans, M. J., “On limits of wireless communications in a fading environment when using multiple antennas,” *Wireless Communication Magazine*, vol. 6, no. 3, pp. 311–335, 1998.
34. DeFlaviis, F., Jofre, L., Jordi, R., Romeu, J., Grau, A., Balanis, C., *Multiantenna Systems for MIMO Communications*, Morgan & Claypool Publishers, 2008.
35. Kuhn, V., *Wireless Communications Over MIMO Channels: Applications to CDMA and Multiple Antenna Systems*, Wiley, New York, 2006.
36. Giannakis, G. B., Liu, Z., Ma, X., Zhou, S., *Space-Time Coding for Broadband Wireless Communications*, Wiley, New York, 2006.
37. Chuah, C.-N., Kahn, J. M., Tse, D., “Capacity of multi-antenna array systems in indoor wireless environment,” *IEEE Globecom*, 1998.
38. Silverstein, J., “Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices,” *Journal of Multivariate Analysis*, vol. 55, no. 2, pp. 331–339, 1995.

Chapter 7
SC-FDMA

7.1 Introduction

Single carrier-frequency division multiple access (SC-FDMA) is an OFDMA variant technology that is tailored for uplink transmission. It uses the standard OFDMA transceiver blocks with different ordering. SC-FDMA (aka DFT-precoded/spread OFDMA) is the multiuser version of single carrier modulation with frequency domain equalization (SC/FDE).

The main objective of SC-FDMA is to introduce transmission with lower PAPR than OFDMA. Since OFDMA shows envelope fluctuations, and signals with high PAPR requires highly linear power amplifiers to reduce the distortion, the design of mobile terminals are complex and they become power hungry since the linearity in the amplifier can only be handled with a large backoff from their peak power.

Another objective is to address frequency offset drawback of OFDMA. In uplink, there are multiple simultaneous transmissions from different mobile stations. If there is slight frequency offset, orthogonality of subcarriers in OFDMA can be destroyed easily.

These issues are addressed in SC-FDMA as follows: unlike OFDMA, which uses parallel transmission, SC-FDMA transmits symbols sequentially so that the PAPR is reduced by spreading a symbol power over subcarriers. Also, SC-FDMA in one mode introduces localized scheduling in which contiguous subcarriers are assigned to a user. This makes mobile station more robust to frequency offset than OFDMA, but of course, the diversity order becomes lower than OFDMA.

Let us first recall OFDMA as explained in detail in the previous chapters and then differentiate toward SC-FDMA.

7.2 SC-FDMA vs. OFDMA

OFDMA utilizes narrow-band orthogonal subcarriers and creates multiple data streams. The transmission rate in each subcarrier is inversely proportional to the total number of orthogonal subcarriers. Number of subcarriers (M) depends on the

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_7, © Springer Science+Business Media LLC 2009

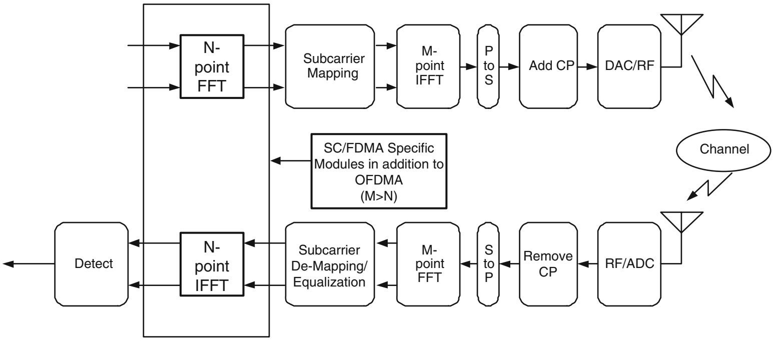


Fig. 7.1 Transmitter and receiver structure for SC-FDMA. *CP* cyclic prefix, *PS* pulse shaping, $M > N$ when SC-FDMA specific module is removed, the structure converges to OFDMA transmitter and receiver

available bandwidth and could be 512, 1,024, or more. As a result, OFDMA system transmits information on M orthogonal subcarriers, each operating bit rate of $1/M$-fold bit rate of the original signal. This rate decrease helps to alleviate the multipath effect of the channel and reduces the equalizer complexity in the receiver. On the other hand, OFDMA suffers from high peak-to-average-power ratio (PAPR). This is due to unpredictable envelope fluctuations after IFFT.

SC-FDMA spreads the energy of one subcarrier over all subcarriers before the IFFT. This way spectral nulls in the channel is reduced with averaging. Hence, PAPR is reduced. This subtle idea is performed by introducing additional FFT block before the IFFT block of the transmitter as seen in Fig. 7.1.

In OFDMA, first, information bits are converted to complex numbers with modulation. Then, the complex numbers are mapped to IFFT block of length M where each number stream is transmitted in a subcarrier out of M. This could be seen as an independent transmission block, and each block produces a time domain signal that are transmitted simultaneously. IFFT block performs these steps and converts these different signal streams from frequency domain into a time domain signal. In uplink OFDMA, of course, each mobile station only uses n subcarriers out of M and leaves the rest null in IFFT process.

In SC-FDMA, these complex numbers are first sent to additional N-point FFT block in order to spread the energy over all the subcarriers. We know that FFT multiplies each complex number with a multiplier and introduces N complex numbers. As a result, output of FFT block is considered as modified complex numbers, and each output contains a portion of every input number. These new modified numbers are sent to M-point IFFT block as in OFDMA. Note that $N < M$ and as in OFDMA, zeros are sent in the unoccupied subcarriers.

In the receiver side, OFDMA utilizes a simple equalizer per subcarrier after FFT. But, SC-FDMA utilizes a complex equalizer before sending the resultant to IFFT. IFFT removes the effect of the FFT in the transmitter. Notice that result of the IFFT is again a time domain signal; the time domain signal is sent to a single detector to create the bits. These differences in receiver side are illustrated in Fig. 7.2,

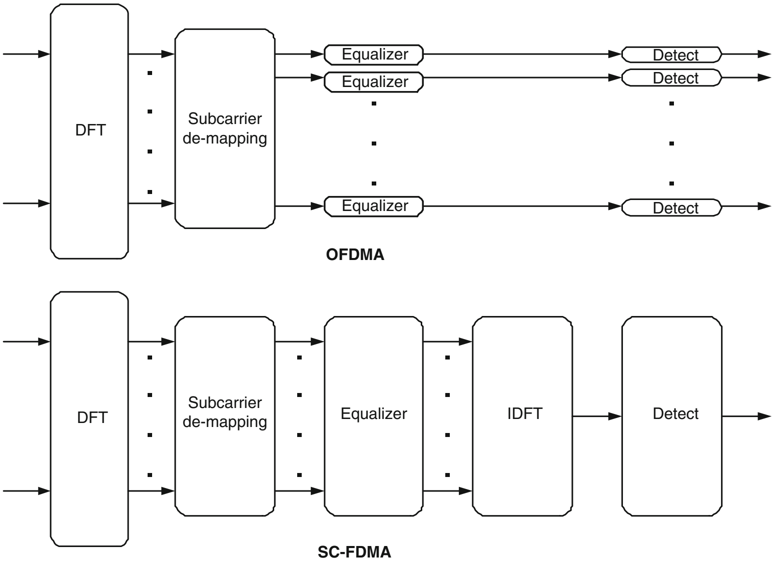


Fig. 7.2 Equalizer comparison in SC-FDMA and OFDMA

where we can see the equalizer simplicity of OFDMA against SC-FDMA. As you can see, SC-FDMA receiver is more complex than OFDMA, but in the transmitter simpler power amplifiers can be utilized to reduce the power consumption. These fortify the SC-FDMA as an uplink transmission scheme, since power efficiency and complexity is important for mobile stations but not in the base station.

7.3 SC-FDMA System

Let us introduce Fig. 7.1 as an uplink SC-FDMA structure to analyze PAPR and resource allocation. Data symbols $\{b_i\}$ are modulated into complex numbers $\{x_i\}$, which are are sent over to N-point FFT system. N-point FFT produces a frequency domain representation (X_n) of the input. After this, each of the parallel output of FFT is sent to a subcarrier of IFFT for transmission resulting $\hat{X}_k$. M-point IFFT transforms $\hat{X}_m$ into time domain complex signals $\hat{x}_m$.

N-point to M-point matching is a resource allocation problem, since $N < M$. $Q = M/N$ is an integer and indicates the number of simultaneous users without any interference since number of users can be increased above M/N with expense on co-channel interference.

Before transmission, first the CP is added to $\hat{x}_m$ and then it is serialized. After that it is modulated with a single frequency carrier. In the receiver side, the received signal is converted to digital format and CP is removed before converting the signal

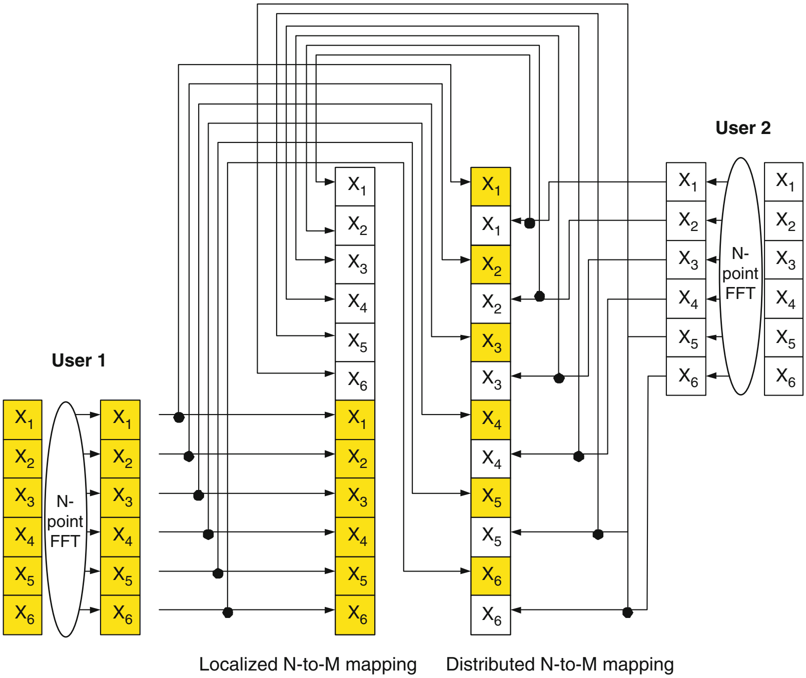


Fig. 7.3 Subcarrier mapping: localized and distributed for two users, where $N = 6$, $M = 12$

into frequency domain with M-point FFT. After channel estimation and equalization, the symbols are sent to N-point IFFT block. Output of the block is sent to the detector to estimate x_i.

The N subcarriers of the user into M subcarriers is mapped in either distributed or localized manner. Figure 7.3 shows an example for the distributed and localized interleaving techniques for two nonoverlapping users.

Distributed mapping (aka interleaved FDMA or IFDMA) introduces bandwidth spreading factor to introduce a parameter for interleaving the allocated subcarriers of a user. IFDMA time sample $\hat{x}_m$ is equal to $\frac{1}{Q}x_{\bar{m}}$ with $\bar{m} = m \bmod N$ if mapping starts from the first subcarrier. Otherwise if mapping starts from r^{th} subcarrier, which is in between 0 to Q, then $\hat{x}_m$ is equal to

$$\hat{x}_m = \frac{1}{Q} e^{j2\pi z(r)} x_{\bar{m}}, \tag{7.1}$$

where $z(r)$ is an additional phase rotation.

Localized mapping (aka localized FDMA or LFDMA) maps subcarriers allocated to user adjacent to each other. LFDMA time samples $\hat{x}_m$ for $r = 0$ is again $\frac{1}{Q}x_{\bar{m}}$, and if $r \neq 0$, then $\hat{x}_m$ equals to

$$\hat{x}_m = \frac{1}{Q.N}\left(1 - e^{j2\pi y(r)}\right) \sum_{i=0}^{N-1} \frac{x_i}{1 - e^{j2\pi w(r,i)}}, \tag{7.2}$$

where $y(r)$ and $w(r,i)$ are additional phase and complex-weighting factors respectively. Notice that there is a $\frac{1}{Q}$ factor in all cases, which basically rounds off the peak power.

IFDMA exploits frequency diversity, since interleaving channel variations can be averaged out. LFDMA on the other hand can be utilized to exploit multiuser diversity, since block of subcarriers can be selected per user according to the channel characteristic. Also note that channel-dependent scheduling does not reach as much diversity order as in OFDMA, since in OFDMA, best subcarriers are selected for a user, but in SC-FDMA, best block of subcarriers is selected, which may not be best for each individual subcarrier within the block. LFDMA also shows larger peak fluctuations in the time domain as compared with IFDMA. This is due to the fact that each user's input may differ from the others and may cause uneven distribution of input symbols. However, in IFDMA, distribution is uniform because of blended inputs. However, frequency synchronization needs to be tighter in IFDMA as compared with LFDMA, thereby LFDMA preserves orthogonality of subcarriers with less complexity. According to these features, IFDMA suits best for high mobile environment, on the other hand LFDMA is good for low mobile environment with channel-dependent scheduling.

Performance of LFDMA and IFDMA peak power shows[1] that IFDMA and LFDMA show lower PAPR than OFDMA. IFDMA is the lowest peak power observed as seen in Fig. 7.4a, and roll-off factor of the raised-cosine pulse shaping filter is inversely proportional to the instantaneous peak power as seen in Fig. 7.4b. The peak power characteristic of LFDMA on the other hand changes with block size as as seen in Fig. 7.4c for a given cut-off w.

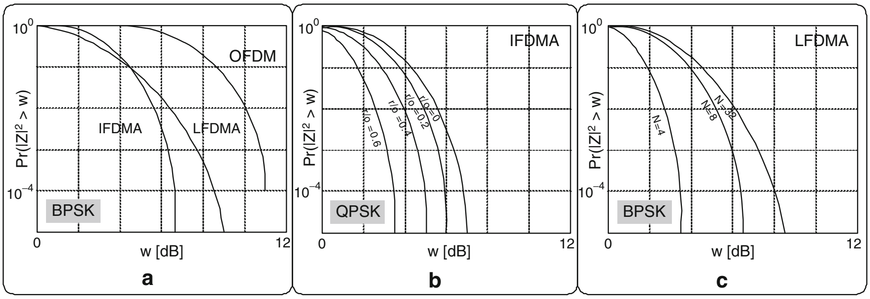


Fig. 7.4 PAPR analysis for upper bound CCDF of SC-FDMA: The distribution of $|x(t)|^2$ of signal $x(t)$ is given with a cut-off filter w. $Pr\{|x(t)|^2 \geq w\}$ is referred as complementary cumulative distribution function (CCDF) and $Z \triangleq x(t_0,\bar{s})$ is a random variable for a given $t_0 \in [0,T)$ and $x(t_0,\bar{s})$ is a baseband representation of the signal carrier modulated signal. $\{s_i\}_{i=-\infty}^{\infty}$ are mutually independent transmitted symbols

[1] "Single Carrier Orthogonal Multiple Access Technique for Broadband Wireless Communications" by Myung, submitted to Electrical and Computer Engineering Department of Polytechnic University, NY, for the degree of Doctor of Philosophy.

7.4 Summary

SC-FDMA is a promising OFDMA-based multicarrier digital technology for uplink. SC-FDMA is being considered to simplify the transmitter in the handsets and reduce the power consumption with lower PAPR feature. In the receiver structure, it is more complex than OFDMA for similar link performance, but this might not be an issue since the receiver is in the base station, which does not have power or complexity limit.

Localized SC-FDMA is considered for LTE uplink against distributed SC-FDMA and OFDMA. Distributed SC-FDMA has not been selected because of its vulnerability to Doppler and frequency offset and its limitation to pilot design. SC-FDMA pilot is generally time-multiplexed with data and designed for low PAPR. This restriction on pilot design is more severe in distributed SC-FDMA and may result in lower flexibility than OFDMA.

SC-FDMA is also proposed to be included to IEEE 802.16m (WiMAX-m) for uplink. But, recent proposals in IEEE 802.16m support OFDMA against SC-FDMA in uplink stating that localized SC-FDMA cannot exploit full advantage of multiuser diversity as in OFDMA and PAPR advantage of SC-FDMA can be mitigated with advanced PAPR techniques in OFDMA. Also backward compatibility to WiMAX-e is another concern when selecting SC-FDMA in uplink of WiMAX-m.

MIMO techniques can be used in SC-FDMA to exploit diversity as well as spatial multiplexing, somewhat similar to MIMO-OFDM, in the frequency domain after FFT as described in the previous chapter.

References

1. Myung, H. G., *Single Carrier Orthogonal Multiple Access Technique for Broadband Wireless Communications*, PhD Dissertation, Polytechnic University, NY, January 2007.
2. Sorger, U., De Broeck, I., Schnell, M., “Interleaved FDMA – A New Spread-Spectrum Multiple-Access Scheme,” *Proceedings of IEEE ICC*, pp. 1013–1017, 1998.
3. Falconer, D., Ariyavisitakul, S. L., Benyamin-Seeyar, A., Eidson, B., “Frequency Domain Equalization for Single-Carrier Broadband Wireless Systems,” *IEEE Communication Magazine*, vol. 40, no. 4, pp. 58–66, 2002.
4. Goodman, D. J., Lim, J., Myung, H. G., “Single Carrier FDMA (SC-FDMA) for Uplink Wireless Transmission,” *IEEE Vehicular Technology Magazine*, 2006.
5. Goodman, D. J., Lim, J., Myung, H. G., “Peak-to-average Power Ratio of Single Carrier FDMA Signals with Pulse Shaping,” *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–5, 2006.
6. Goodman, D. J., Lim, J., Myung, H. G., Oh, K., “Channel-Dependent Scheduling of Uplink Single Carrier FDMA Systems,” *Proceedings of IEEE VTC*, 2006.
7. Batariere, M. D., Classon, B. K., “Low-Complexity Technique to Increase Capacity of Mobile Broadband Systems,” *Proceedings of IEEE VTC*, vol. 4, pp. 1939–1943, 2000.
8. Cioffi, J. M., Tellado, J., “PAR Reduction in Multicarrier Transmission Systems,” *ANSI* T1E1.4/97-367, 1997.

9. Lopez, P., Monnier, R., Tourtier, P. J., “Multicarrier Modem for Digital HDTV Terrestrial Broadcasting,” *Signal Processing: Image Communication*, vol. 5, no. 6, pp. 379–403, 1998.
10. VDSL Alliance, *VDSL Alliance SDMT VDSL Draft Standard Proposal*, ANSI Contribution T1E1.4/97-332, 1997.
11. Leung, C., Warner, W. D., “OFDM/FM Frame Synchronization for Mobile Radio Data Communication,” *IEEE Transactions on Vehicular Technology*, vol. 42, no. 3, pp. 302–313, 1993.

Chapter 8
WiMAX Physical Layer

WiMAX air interface is based on a IEEE standard[1] emerged from 802.16 working group. The IEEE 802.16 Working Group focuses on Broadband Wireless Access Standards. The group develops standards and recommended practices to support the development and deployment of broadband Wireless Metropolitan Area Networks (Wireless MAN).

IEEE 802.16-2004 and IEEE 802.16e-2005 standards[2] construct the basis of Mobile WiMAX Physical (PHY) layer and Medium Access Layer (MAC). 802.16 series defines four Wireless MAN PHY layers and any of them can be combined with the MAC layer, described in the next chapter;

- **WirelessMAN-SC:** WirelessMAN-SC is the first standard that is introduced by 802.16 working group. It employs a single-carrier (SC) line-of-sight (LOS) modulation for point-to-point communication to operate in the 10–66 GHz spectrum. This standard is to address network access support to buildings with data rates that is comparable to those offered by high-speed fiber optic networks.
- **WirelessMAN-SCa:** LOS communication in SC is ratified with 802.16a-2003 amendment to address low-frequency 2–11 GHz spectrum with non-line-of-sight (NLOS) point-to-multipoint communication for fixed broadband wireless access. Later, SCa is again ratified with 802.16d in 2003 and finalized in 802.16-2004.
- **WirelessMAN-OFDM:** 802.16a-2003 added an OFDM PHY with 256 subcarriers to accommodate NLOS fixed access for frequencies in 2–11 GHz. Later, it is finalized in 802.16-2004 standard. This is the approved WiMAX fixed access standard by WiMAX Forum.
- **WirelessHUMAN:** Wireless High-Speed Unlicensed MAN is similar to OFDM PHY but dynamic frequency selection is mandatory for license exempt bands.

[1] "People often take the view that standardization is the enemy of creativity. But I think that standards help make creativity possible – by allowing for the establishment of an infrastructure, which then leads to enormous entrepreneurialism, creativity, and competitiveness." by Vint Cerf, commonly referred as one of the *founding fathers of the Internet*."

[2] Approved WiMAX mobile and portable access standard by WiMAX Forum.

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_8, © Springer Science+Business Media LLC 2009

- **WirelessMAN-OFDMA:** 802.16a-2003 also introduced up to 2,048-carrier OFDMA PHY to accommodate NLOS point-to-multipoint communication. This is ratified in 802.16-2004 and revisited in 802.16e-2005 for mobile access. OFDMA delivers QoS at high speeds with the following features:
  - Scalable channel bandwidths from 1.25 to 20 MHz[3] with Fast Fourier Transform (FFT) size.
  - Resistant to interference with subchannel orthogonality and reduction of Inter-carrier and Inter-symbol interferences with guard band.
  - Time Division Duplex (TDD) support for asynchronous data traffic and implicit channel side information via channel reciprocity.
  - Significant cell range extension due to concentrated transmit power in uplink and with assignment of more power to distant users in downlink.
  - Hybrid-Automatic Repeat Request (HARQ) reliability support for high mobile situations.
  - Frequency selective scheduling and subchannelization with various flexible permutation options.
  - Sleep and idle mode support for power management.
  - Optimized hard handover and support for soft handover.
  - Multicast and Broadcast Service with the inherited features of DVB-H, MediaFLO, and 3GPP E-UTRAN for high data rate coverage in single frequency network.

Additionally, 802.16 group focuses on following ongoing and completed amendments:

- **802.16m** is currently in predraft stage and being designed to focus on advanced air interface to meet the cellular layer requirements of IMT-Advanced next generation mobile networks. It is an amendment to air interface for fixed and mobile broadband wireless access services to push data rates up to 100 Mbps for mobile and 1 Gbps for fixed while maintaining backward compatibility with existing WiMAX radios. The 802.11 m is designed to fully utilize MIMO technology on top of an OFDMA based radio system just like the upcoming "Wave 2" mobile WiMAX products which use two-by-two antenna to achieve mobile speeds of around 5 Mbps. "Wave 3" mobile WiMAX, on the other hand, will consider four-by-four smart antenna array.
- **802.16h** is in draft stage and being designed to focus on improving coexistence mechanisms for license-exempt operation as an amendment to air interface for fixed and mobile broadband wireless access systems. The goal is to ensure that multi-vendor WiMAX systems can be readily deployed in the license exempt bands with regard to minimum interference to other deployed 802.16 based license exempt deployment.
- **802.16i** is in draft stage and being designed to focus on mobile management information base for MAC, PHY, and associated management procedures. The aim of the standard is to develop protocol-neutral methodologies for network management for multi-vendor operation.

[3] WiMAX Forum Release 1 supports 5, 7, 8.75, and 10 MHz.

- **802.16j** is in draft stage and being designed to focus on providing multihop relay specification as an amendment to air interface for fixed and mobile broadband wireless access systems. The standard specifies OFDMA PHY and MAC enhancement to enable the operation of relay stations in licensed bands. The purpose is to enhance coverage, throughput and system capacity of 802.16 networks with interoperable relay and base stations.
- **802.16g** is an active standard and being designed to focus on providing management plane procedures and services as an amendment to air interface for broadband wireless access systems. The purpose of the amendment is to provide conformant 802.16 equipment with procedures and services to enable interoperable and efficient management of network resources, mobility, and spectrum.
- **802.16f** is an active standard and being designed to focus on providing management information base as an amendment to air interface for fixed broadband wireless access systems.
- **802.16k** is published standard and designed to focus on bridging of 802.16 as media access control bridges for local and metropolitan area networks.

In this chapter, we cover mobile WiMAX physical layer (IEEE 802.16e-2005 OFDMA mode) in detail with emphasis on MIMO support. First, we talk about constructing the OFDMA signal, a OFDMA signal is made up of subcarriers in frequency domain. Then, we discuss the underlying foundation of slot structure, which is constructed basically grouping the subcarriers in several mathematical fashion to exploit transmit or multiuser diversity. After this, we talk about frame structure, which is composed of slots that spans in time and frequency domain. An OFDMA frame can address many subscribers[4] and certain configurations like adaptive modulation, zoning are possible to accommodate subscribers with different characteristics. Later, we give the foundation of open and closed-loop space-time coding and MIMO, supported in the standard. This introduction is followed by coding mechanism including HARQ and supported control mechanisms such as ranging, power control, and channel quality measurements.

8.1 OFDMA Signal

WiMAX introduces a scalable OFDMA concept, termed "Scalable OFDMA" or "SOFDMA" where number of used subcarriers N_{used} scales with channel bandwidth BW. This way subcarrier spacing Δf is kept constant for all bandwidths and determined by $F_{\text{s}}/N_{\text{FFT}}$ where N_{FFT} is smallest power of two greater than N_{used} and $F_{\text{s}}(= floor(n.BW/8000) \times 8000)$, sampling frequency, is determined with sampling factor n. This value is set to $n = 8/7$ as default but if channel bandwidths are multiple of 1.25, 1.5, 2, or 2.75 MHz, then $n = 28/25$.

[4] Throughout this book, the terms subscriber, mobile station, mobile subscriber, and WiMAX subscriber are used interchangeably to refer to the client device that communicates to the base station through air interface.

OFDMA useful symbol time T_b is $1/\Delta f$ and it is extended with cyclic guard interval $T_\mathrm{g} = G.T_\mathrm{b}$, termed CP. Typically, subscriber searches for CP of base station and uses that CP on the uplink. Guard ratio (G) to "useful" time could be 1/32, 1/16, 1/8, and 1/4. As a result, OFDMA symbol time becomes $T_\mathrm{s} = T_\mathrm{b} + T_\mathrm{g}$ with sampling time of $T_\mathrm{b}/N_\mathrm{FFT}$. Finally, transmitted signal to the antenna as a function of time is given by

$$s(t) = Re\left\{ \mathrm{e}^{j2\pi f_\mathrm{c}t} \sum_{\substack{k=-(N_\mathrm{used}-1)/2 \\ k \neq 0}}^{(N_\mathrm{used}-1)/2} c_k.\mathrm{e}^{j2\pi k\Delta f(t-T_\mathrm{g})} \right\}, \tag{8.1}$$

where t is the time ($0 < t < T_\mathrm{s}$) and c_k is a complex number, which is basically the data to be transmitted on the subcarrier whose frequency offset index is k, during the subject OFDMA symbol. It basically specifies a point in a QAM constellation.

8.2 OFDMA Symbol

As we said, symbols are made up of subcarriers. Total number of subcarriers (N_FFT) are made up of either data subcarriers for data transmission, or pilot subcarriers for estimation purposes, or null carriers for no transmission. Otherwise, it is either used for guard bands or a DC carrier.

Active subcarriers are divided into subsets of subcarriers, each subset is termed a subchannel with either *distributed* subcarrier permutation or *adjacent* subcarrier permutation. This subchannelization allows scalability, multiple access, and advanced antenna array processing capabilities. We introduce subcarrier to subchannel mapping for adjacent and distributed subcarrier permutations in this section. Further, subchannels are clustered into six subchannel groups to provide flexible allocation and segmentation that we address in the next section.

8.2.1 FUSC: Full Usage of Subcarriers

FUSC method applies distributed subcarrier permutation by first allocating the pilot subcarriers and then terming each remaining subcarrier to a subchannel. There are 48 subcarriers per subchannel. A pilot subcarrier can either be from constant set or variable set. The location of pilot subcarriers of constant set is fixed but location of pilot subcarriers from variable set changes with each symbol. This is to accurately estimate channel with large delay spread. Figure 8.1 shows the process to create the subchannels from subcarriers with parameters indicated in Table 8.1. According to a permutation formula, data subcarriers are partitioned into groups of contiguous

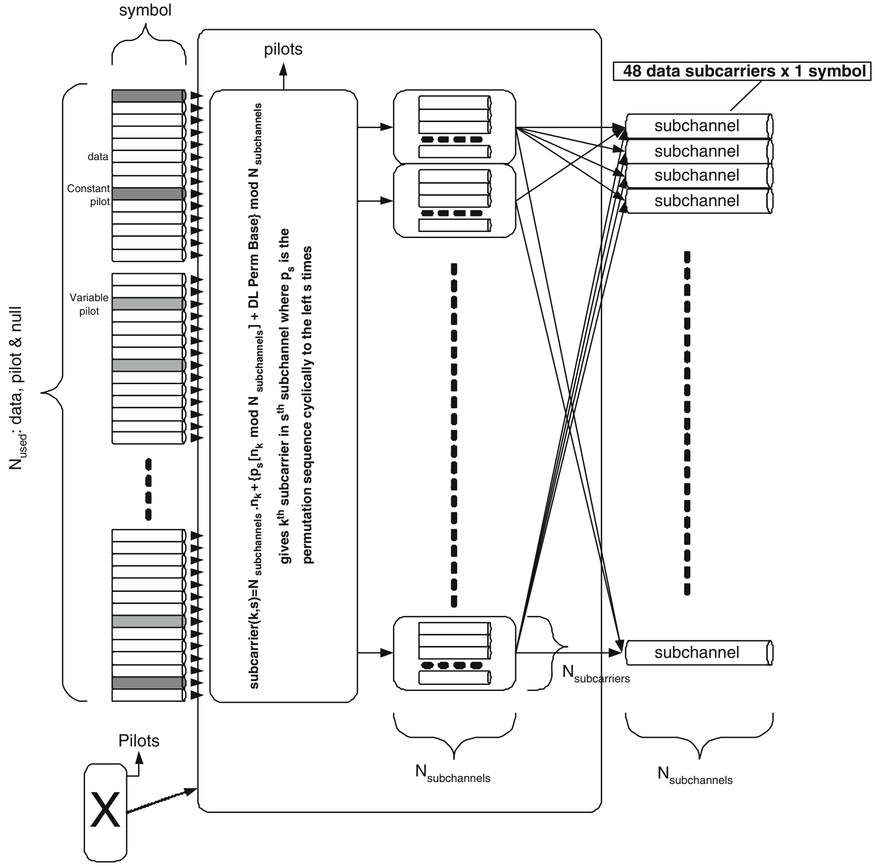

symbol
pilots
48 data subcarriers x 1 symbol
data
Constant pilot
Variable pilot
subcarrier(k,s)=N subchannels ·n_k+{p_s[n_k mod N subchannels] + DL Perm Base} mod N subchannels
gives k^th subcarrier in s^th subchannel where p_s is the
permutation sequence cyclically to the left s times
subchannel
subchannel
subchannel
subchannel
subchannel
N_used: data, pilot & null
N_subcarriers
N_subchannels
N_subchannels
Pilots
X

Fig. 8.1 FUSC

Table 8.1 OFDMA downlink carrier allocations - FUSC

Parameter/FFT size	2,048	1,024	512	128
N_{DC}	1	1	1	1
$N_{\text{guard}}, \mathit{left}$	173	87	43	11
$N_{\text{guard}}, \mathit{right}$	172	86	42	10
N_{used}	1,703	851	427	107
N_{pilot}	–	–	–	4
$N_{\text{variable}}/N_{\text{constant}}, \mathit{even}$	71/12	12/2	18/3	5/1
$N_{\text{variable}}/N_{\text{constant}}, \mathit{odd}$	71/12	12/2	18/3	4/0
N_{data}	1,536	768	384	96
$N_{\text{data/subchannel}}$	48	48	48	48
$N_{\text{subchannels}}$	32	16	8	2

Table 8.2 OFDMA downlink carrier allocations - optional FUSC

Parameter/FFT size	2,048	1,024	512	128
N_{DC}	1	1	1	1
$N_{\text{guard}}, left$	160	80	40	10
$N_{\text{guard}}, right$	159	79	39	9
N_{used}	1,729	865	433	109
N_{pilot}	192	96	48	12
N_{data}	1,536	768	384	96
$N_{\text{data/subchannel}}$	48	48	48	48
$N_{\text{subchannels}}$	32	16	8	2

subcarriers. Each subchannel consists of one subcarrier from each of these groups. The number of groups is therefore $N_{\text{subchannels}}$ and number of subcarriers in each group is $N_{\text{subcarriers}}$.

Additional optional subchannel structure for FUSC defines new formula for pilot allocation in which one pilot is allocated to nine contiguous subcarriers according to following formula:

$$\text{pilot\_index} = 9k + 3(\text{symbol\_index})_{\text{mod }3} + 1, \tag{8.2}$$

where k is the subcarrier index and symbol index starts from 0. The data subcarriers are partitioned into $N_{\text{subchannels}}$ groups and a subchannel consists of one subcarrier from each of these groups. Parameters for optional subchannel structure for downlink FUSC are presented in Table 8.2.

8.2.2 DL PUSC: Downlink Partial Usage of Subcarriers

DL PUSC is also distributed permutation scheme, which first divides the subcarriers into N_{clusters} clusters, each holding 14 adjacent subcarriers over two symbols. Clusters are numbered with pseudorandom numbering together with a permutation base metric and then allocated into six groups. In a cluster, there are 28 subcarriers and 24 of them are used for data and 4 of them are used for pilot subcarriers. Allocating subcarriers to subchannels in each major group is performed separately for each OFDMA symbol by first allocating the pilots and then taking all data subcarriers within the symbol to term them to a subchannel as in FUSC. Permutation sequences for even and odd symbol are different and 24 subcarriers are allocated to a subchannel in each symbol. Table 8.3 shows the parameters for FFT sizes 2048, 1024, 512, and 128 where PUSC procedure for downlink is illustrated in Fig. 8.2.

8.2.3 UL PUSC: Uplink Partial Usage of Subcarriers

UL PUSC is a little different, first usable subcarriers are divided into N_{tiles}. These tiles are renumbered with a pseudorandom scheme. A tile hosts 8 data and 4 pilot

Table 8.3 OFDMA downlink carrier allocations - PUSC

Parameter/FFT size	2,048	1,024	512	128
N_{DC}	1	1	1	1
N_{guard}, *left*	184	92	46	22
N_{guard}, *right*	183	91	45	21
N_{used}	1,681	841	421	85
$N_{\text{subcarrier/cluster}}$	14	14	14	14
N_{clusters}	120	60	30	6
$N_{\text{data/symbol/subchannel}}$	24	24	24	24
$N_{\text{subchannels}}$	60	30	15	3

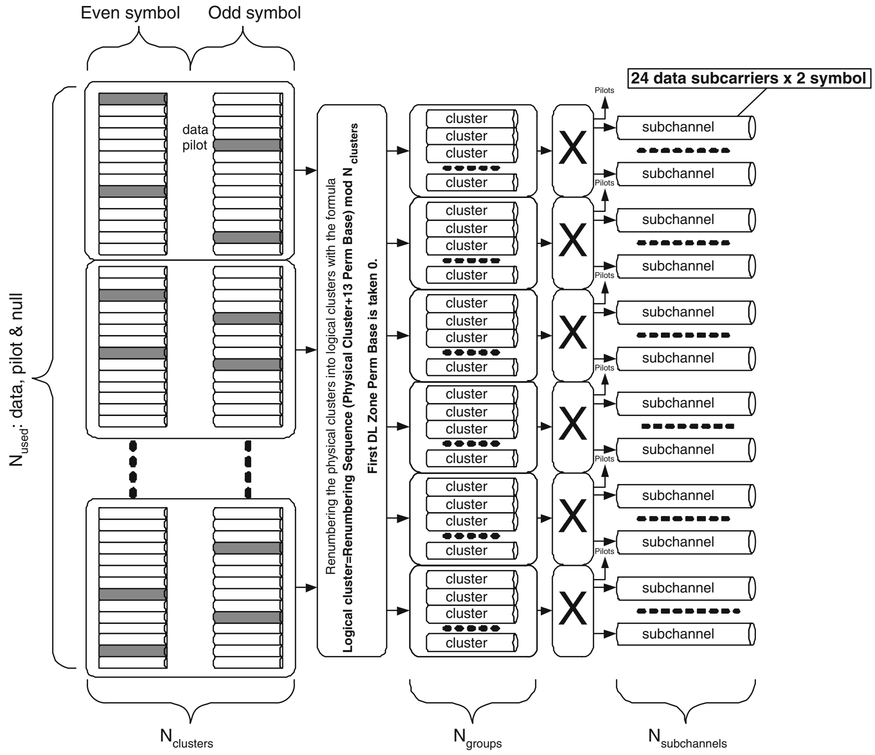

Even symbol
Odd symbol
data
pilot
N_{used}: data, pilot & null
Renumbering the physical clusters into logical clusters with the formula
Logical cluster=Renumbering Sequence (Physical Cluster+13 Perm Base) mod N_{clusters}
First DL Zone Perm Base is taken 0.
cluster
X
Pilots
24 data subcarriers x 2 symbol
subchannel
N_{clusters}
N_{groups}
$N_{\text{subchannels}}$

Fig. 8.2 DL PUSC

subcarriers over 4 subcarriers to 3 symbols as seen in Fig. 8.3. Pseudorandom scheme renumbers the tiles and creates groups with six tiles. A subchannel is created with six tiles form the same group. There is also optional PUSC mode in which tile structure hosts 8 data and 1 pilot with 3 subcarriers to 3 symbols as seen in Fig. 8.4. The optional PUSC is suitable if higher data rate is preferred over coarse channel estimation. Tables 8.4 and 8.5 show the parameter for UL PUSC and optional UL PUSC.

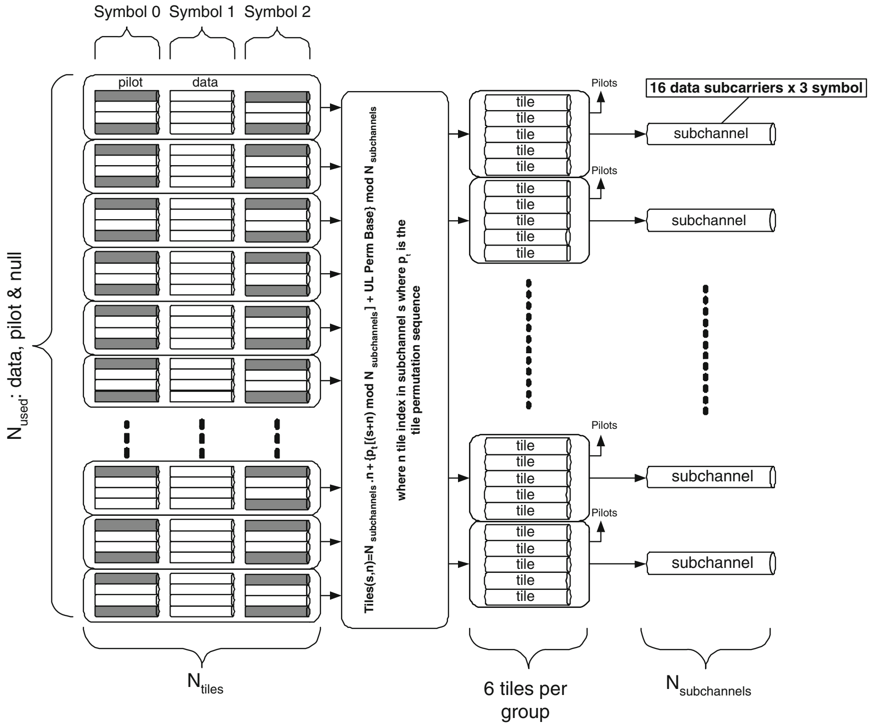

Symbol 0
Symbol 1
Symbol 2
pilot
data
N_{used}: data, pilot & null
N_{tiles}
Tiles(s,n)=N subchannels ·n + {p_t[(s+n) mod N subchannels] + UL Perm Base} mod N subchannels
where n tile index in subchannel s where p_t is the tile permutation sequence
tile
Pilots
16 data subcarriers x 3 symbol
subchannel
6 tiles per group
$N_{subchannels}$

Fig. 8.3 UL PUSC

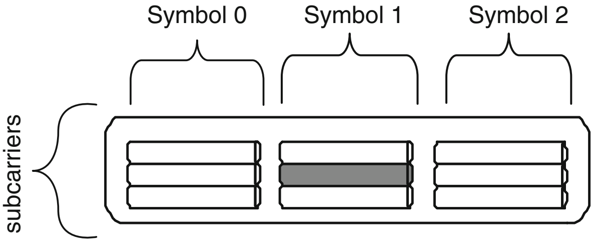


Fig. 8.4 Tile Structure for optional UL PUSC

8.2.4 TUSC: Tile Usage of Subcarriers

Channel reciprocity is required for closed-loop AAS systems. To eliminate explicit channel feedback, downlink and uplink can be similarly configured to achieve channel reciprocity in TDD mode. The TUSC, only permitted in downlink with AAS, has two modes: TUSC1 and TUSC2, which are identical to UL PUSC and optional UL PUSC, respectively.

Table 8.4 OFDMA uplink carrier allocations - UL PUSC

Parameter/FFT size	1,024	512	128
N_{DC}	1	1	1
$N_{\text{guard}}, left$	92	52	16
$N_{\text{guard}}, right$	91	51	15
N_{used}	841	409	97
$N_{\text{subchannels}}$	35	17	4
N_{tiles}	210	102	24
$N_{\text{subcarriers/tile}}$	4	4	4
$N_{\text{tiles/subchannel}}$	6	6	6

Table 8.5 OFDMA uplink carrier allocations - optional PUSC

Parameter/FFT size	2,048	1,024	512	128
N_{DC}	1	1	1	1
$N_{\text{guard}}, left$	160	80	40	10
$N_{\text{guard}}, right$	159	79	39	9
N_{used}	1,729	865	433	109
$N_{\text{subchannels}}$	96	48	24	6
N_{tiles}	576	288	144	36
$N_{\text{subcarriers/tile}}$	3	3	3	3
$N_{\text{tiles/subchannel}}$	6	6	6	6
$N_{\text{subcarriers/subchannel}}$	48	48	48	48

8.2.5 AMC Subchannels

So far, we introduced distributed subcarrier permutation that leverages frequency diversity. Subchannel assignment can be with adjacent subcarrier permutation in which pilot and data subcarriers are assigned fixed positions. This type of permutation scheme is used to leverage multiuser diversity since a subcarrier can be assigned to the user with best channel for that subcarrier. A BS may switch from distributed to adjacent subcarrier permutation for AAS traffic with an indication in the frame.

Figure 8.5 shows the AMC subcarrier permutation where a *bin* is created from 8 data subcarrier and 1 pilot. An AMC subchannel of type $N \times M = 6$ is defined as six contiguous bins where a slot is N bins by M symbols. Figure 8.5 is for 3×2 AMC subchannel and 2×3, 1×6 slot configurations are also possible. Parameters for AMC is depicted in Table 8.6.

8.2.6 Data Rotation

To exploit frequency diversity further, each OFDMA slot can be rotated in any zone except AAS zone, optional PUSC zone, or zone using the adjacent-subcarrier

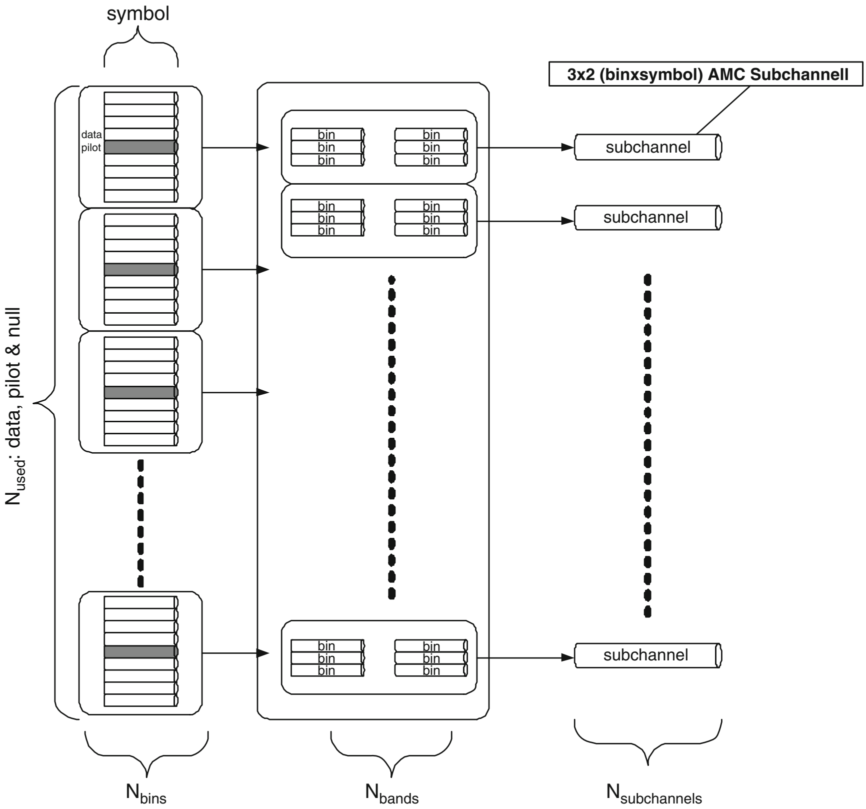


Fig. 8.5 AMC subcarrier permutation

Table 8.6 OFDMA AAS subcarrier allocations - AMC

Parameter/FFT size	2,048	1,024	512	128
N_{DC}	1	1	1	1
N_{guard}, $left$	160	80	40	10
N_{guard}, $right$	159	79	39	9
N_{used}	1,729	865	433	109
N_{pilots}	192	96	48	12
N_{data}	1,536	768	384	96
N_{bands}	48	24	12	3
$N_{bins/bands}$	4	4	4	4
$N_{datasubcarrier/subchannel}$	48	48	48	48

permutations. During slot duration, subchannels are renumbered contiguously and mapping function gives the new subchannel number. This applies to uplink as well for some restricted areas.

8.3 OFDMA Frame

We covered the procedures in which physical subcarriers are mapped to subchannels. Subchannel concept is typically used to introduce the frequency component of a **slot**, which is the minimum possible data allocation unit. An OFDMA slot depends on the OFDMA symbol structure where various slot configurations are seen in Fig. 8.6:

- Downlink FUSC with distributed subcarrier permutation: 1 SLOT = 1 Subchannel by 1 OFDMA Symbol, which is 48 data subcarriers over 1 symbol.
- Downlink PUSC with distributed subcarrier permutation: 1 SLOT = 1 Subchannel by 2 OFDMA Symbol, which is 2 clusters (48 data subcarriers) span over two symbols.
- Uplink PUSC with distributed subcarrier permutation: 1 SLOT = 1 Subchannel by 1 OFDMA Symbol, which is 6 tiles (48 data subcarriers) span over three symbols.

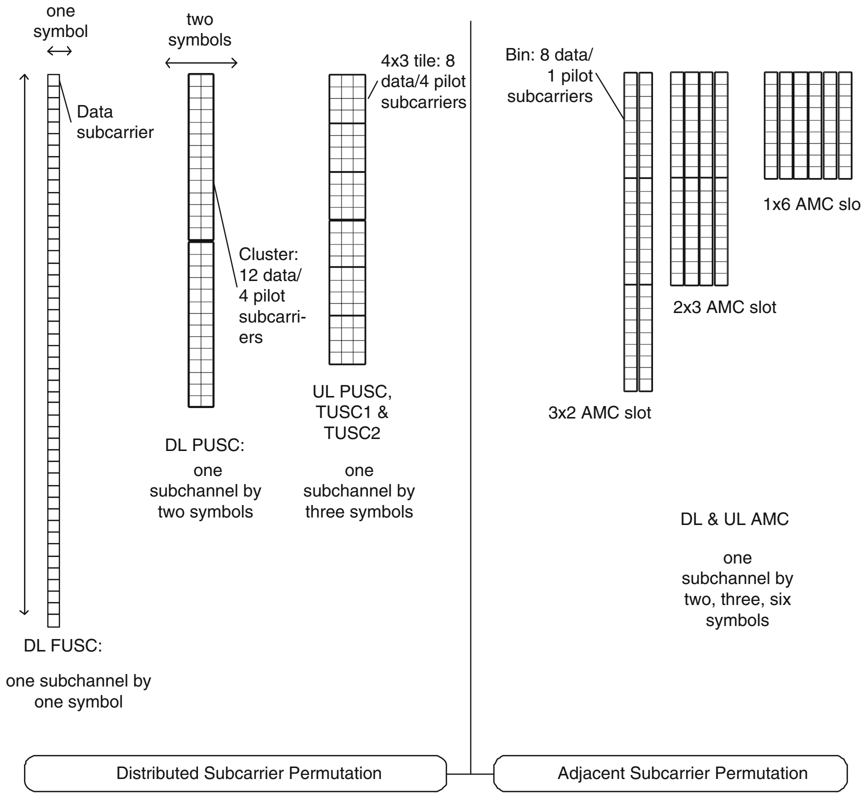

one symbol
two symbols
Data subcarrier
4x3 tile: 8 data/4 pilot subcarriers
Bin: 8 data/ 1 pilot subcarriers
1x6 AMC slot
Cluster: 12 data/ 4 pilot subcarriers
2x3 AMC slot
UL PUSC, TUSC1 & TUSC2
3x2 AMC slot
DL PUSC:
one subchannel by two symbols
one subchannel by three symbols
DL & UL AMC
one subchannel by two, three, six symbols
DL FUSC:
one subchannel by one symbol
Distributed Subcarrier Permutation
Adjacent Subcarrier Permutation

Fig. 8.6 Slot structures

- Uplink and Downlink with adjacent subcarrier permutation: 1 SLOT = 1 Subchannel by 2,3, or 6 OFDMA Symbols, which is M bins span over N symbols, where $M.N = 6$ and total is again 48 data subcarriers.

There are also specific frame durations where once set, it should not be changed. Otherwise, it requires resynchronization of base stations. Available frame durations are $\{2.5, 4, 5, 8, 10, 12.5, 20\,\text{ms}\}$.

8.3.1 OFDMA Data Mapping

Slots positioned in time and frequency constitutes a **frame**.There is a **data region** within a frame defined by subchannel and symbol dimension as seen in Fig. 8.7. Before mapping, data are segmented into blocks, which is sized to fit into one OFDMA slot.

In TDD, frame period is divided into two sections in time: first section is for DL transmission with a preamble; second section is for UL transmission. Between two sections, there are gaps to provide sufficient time to allow BS to turn around.

A data region is addressed by DL-MAPs or UL-MAPs, which are transmitted in the beginning of frame as seen in Fig. 8.8. These MAPs are broadcasted and have entries for all the data regions allocated to any subscriber. An allocation is specified with a MAP Information Element (MAP_IE). MAP_IE includes connection identifier (CID), which is unique identifier per flow within a base station. Also MAP_IE defines the coordinates of the data region in the frame with subchannel offset and symbol duration together with burst profile.

In downlink, mapping starts from the lowest numbered OFDMA subchannel and spans one or more OFDMA symbols depending on the slot definition. Mapping continues in subchannel direction till the edge of the data region is reached. After that, mapping continues with the lowest numbered subchannel in the next available symbol.

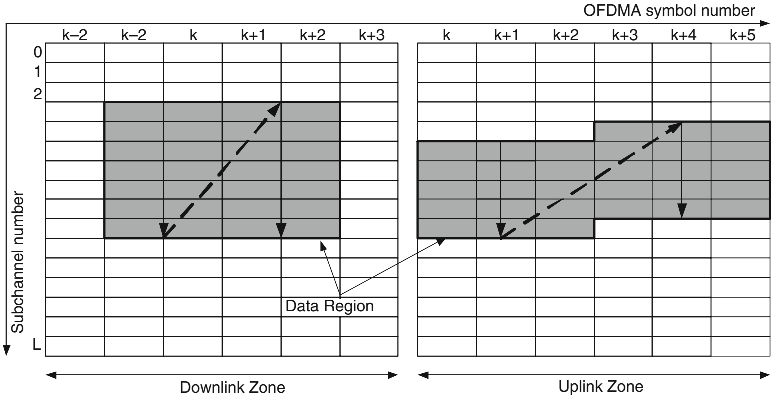


Fig. 8.7 Mapping for PUSC mode

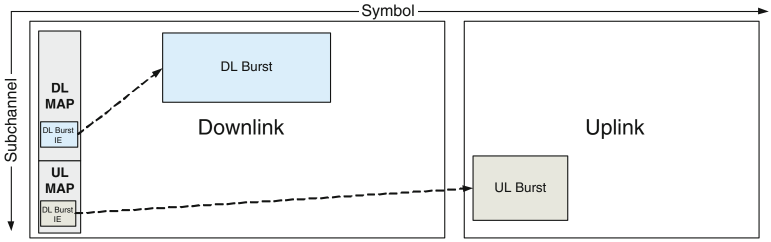


Fig. 8.8 Addressing in a frame: a MAP_IE content includes CID, symbol and subchannel offset, symbol duration and number of subchannels, and burst profile. Maximum number of bursts the MS can decode in one downlink frame is 64 and the maximum number of bursts that can be transmitted concurrently is 16

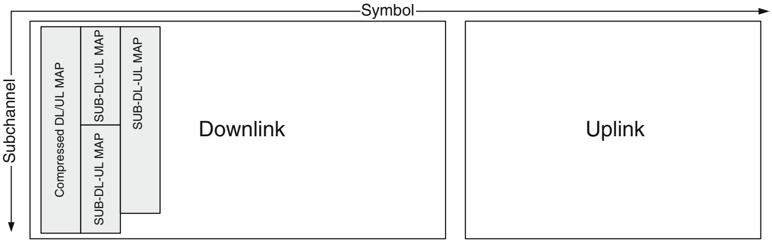


Fig. 8.9 Sub-MAP Layout: Sub-MAP's created to reduce MAP overhead. MAP overhead for bursty data traffic (FTP, HTTP) $\sim$ 10%, MAP overhead for VoIP traffic increases as number of users increase >20%. MAP is sent in lowest modulation

Uplink mapping consists of two steps: first step, the OFDMA slots are allocated to each burst in increasing symbol number; in the second step, the allocated slots are mapped in increasing subchannel number. In UL-MAP, a transportation opportunity is also allocated to allow group of stations to transmit for bandwidth requests or ranging requests for network entry.

MAPs can be compressed and submaps can be created to save bandwidth as seen in Fig. 8.9. Sub-MAPs can be sent at different data rates.

8.3.2 TDD Frame

A typical TDD frame is seen in Fig. 8.10. First, a preamble is sent on all subchannels starting at symbol offset 0. For preamble, subcarriers are modulated with BPSK modulation with a Pseudo-Noise (PN) code. There are specific preamble modulation series per segment with specified cell ID. After preamble there is Frame Control Header (FCH) message, which is located in the first 4 slots of any segment. FCH contains information about the frame format such as MAP length, the repetition coding used in MAP, a bitmap for the subchannels that are in use and ranging information. FCH is sent always with QPSK and a FEC of $R = \frac{1}{2}$ coding rate with four

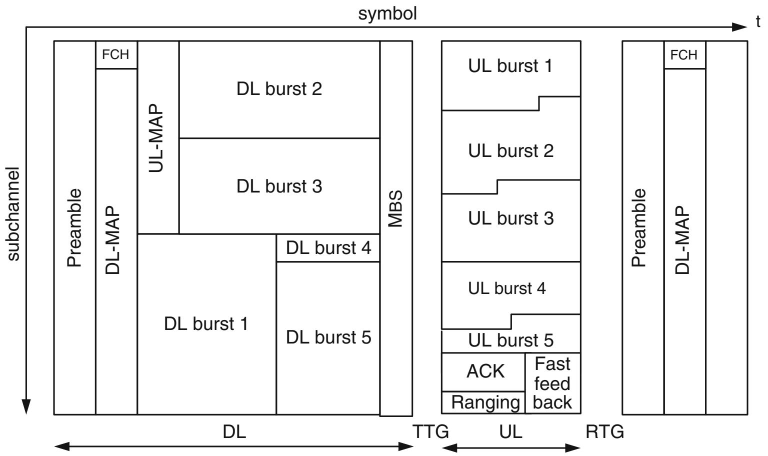


Fig. 8.10 A typical TDD frame

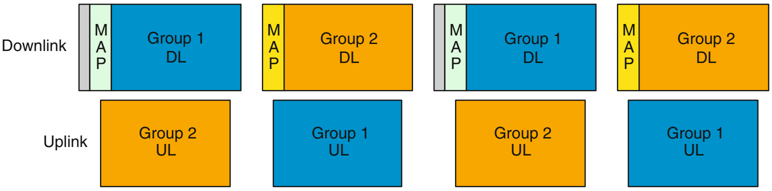


Fig. 8.11 OFDMA FDD/HFDD frame (under consideration)

repetitions. For FFT sizes 2,048, 1,024, 512, maximum MAP length is 256 slots: for FFT size 128, maximum MAP length is 32 slots. The DL-MAP comes in the next slot after the FCH.

In each frame, there are two gaps inserted to allow the BS to turn around: Transmit/Receive Transition Gap (TTG) is a silence zone between DL and UL transmission, Receive/Transmit Transition Gap (RTG) is another silence zone between two frames.

Subscriber station allowances for TDD and Hybrid-FDD (HFDD) systems are *SS*RTG[5] and *SS*TTG. Before its scheduled uplink allocation, the BS transmits downlink information to a station later than (*SS*RTG+RTD). The BS does not transmit downlink information to it earlier than (*SS*TTG-RTD) after the end of scheduled uplink allocation. Notice that RTD denotes Round-Trip Delay and during network entry these parameters are provided (Fig. 8.11).

[5] SS stands for subscriber station.

8.3.3 FDD/HFDD Frame

The Frequency Division Duplex (FDD) (Full-Duplex and Half-Duplex operations) operation support for WiMAX is being designed within WiMAX Forum. The FDD or HFDD utilizes two-way radio in which the transmitter and receiver operate in different frequencies. Full-Duplex operation enables transmission and reception at the same time and Half-Duplex operation only allows transmission or reception at any given time, which requires switching between transmission and reception.

Figure 10.30 depicts a possible frame structure where two virtual groups are introduced. In HFDD operation, MSs that belong to a group are only allowed to transmit and receive at the assigned partition. For FDD operation, a MS may belong to both groups.

8.3.4 Segments and Zones

A frame can be divided into segments in frequency axis and zones in the time axis. In the frequency axis, subchannels are grouped into subchannel groups, which can be clustered further to form **segments** with PUSC as seen in Fig. 8.12. For instance, a MAC instance can be deployed to each segment as seen in Fig. 8.13 to enable single frequency network. Figure 8.14 depicts the deployment strategies with segmentation to increase frequency reuse.

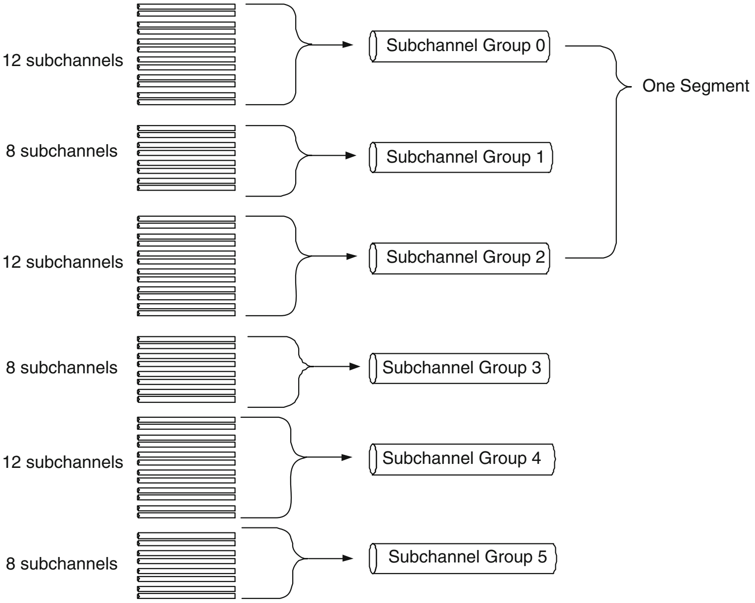


Fig. 8.12 Segment partitioning

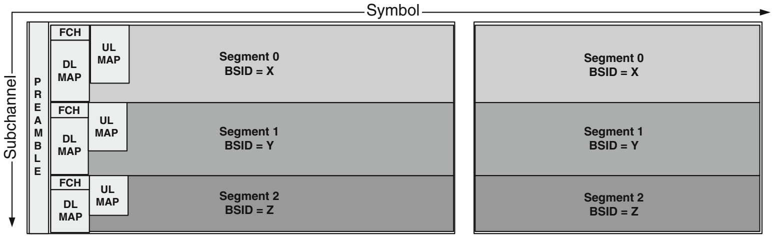


Fig. 8.13 Segment partitioning in frame

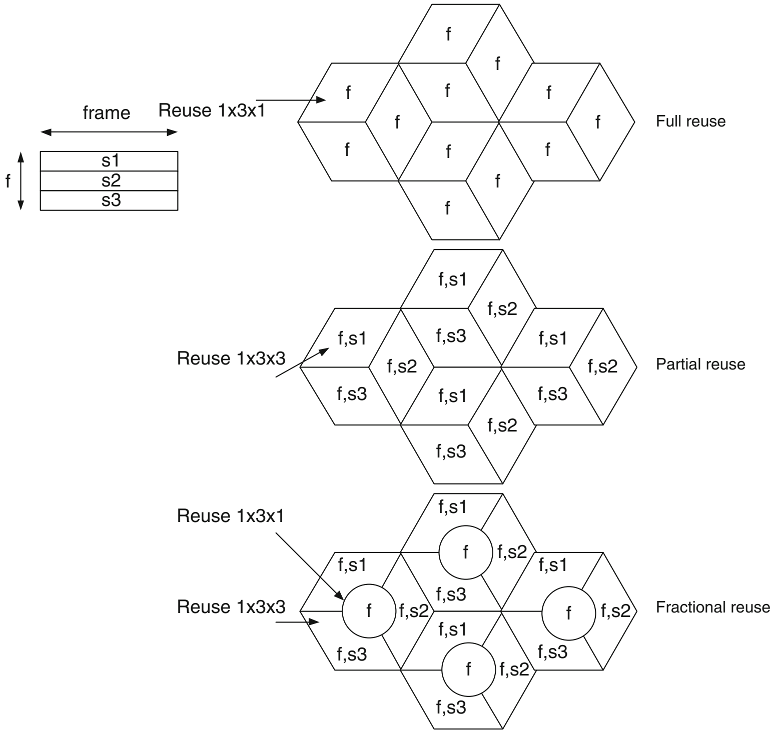


Fig. 8.14 Frequency reuse with segmentation: full reuse as in 1 × 3 × 1 is not possible due to interference in the overlapping regions. Partial reuse as in 1 × 3 × 3 can be increased further with fractional reuse

In the same way, symbols can be clustered into **zones** in symbol dimension, which could have different subcarrier to subchannel mapping formula. This is efficient to accommodate users with different characteristic; for instance, scheduling the high mobile and low mobile users into different zones, etc. The transition between zones is indicated in the MAP as seen in Fig. 8.15. Downlink MAP contains a Zone Switch indicator with a symbol offset and permutation number (DL_Permbase).

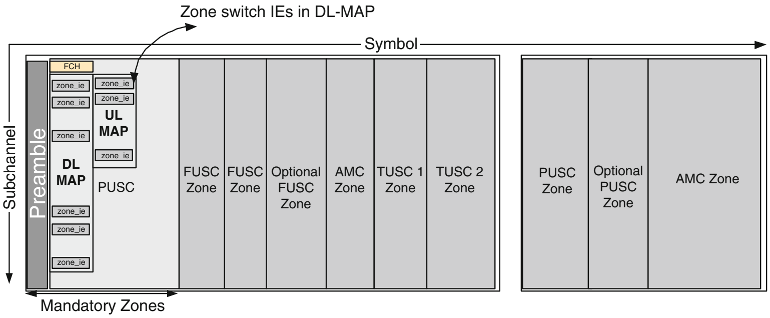


Fig. 8.15 Frame Zone Layout: zoning physically subdivides the downlink and uplink portions of the frame into smaller sections of time, all symbols in a zone use the same permutation formula and maximum 8 downlink zones are supported

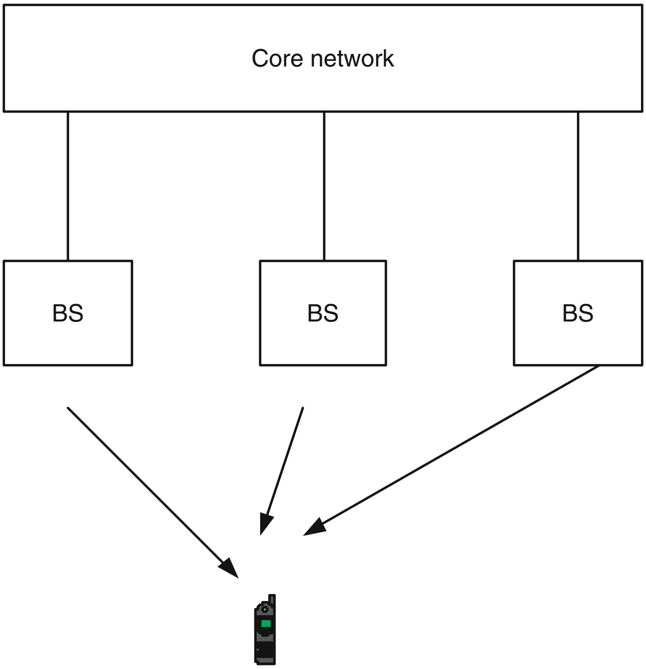


Fig. 8.16 Multicast broadcast service

8.3.5 MBS Zone

Multicast Broadcast Service is a broadcast service to support delivering media content as seen in Fig. 8.16. BSs participating in MBS transmit the same data at the same time and at the same location in a given frame. Stations receiving multiple transmission over the air combine them for better reception.

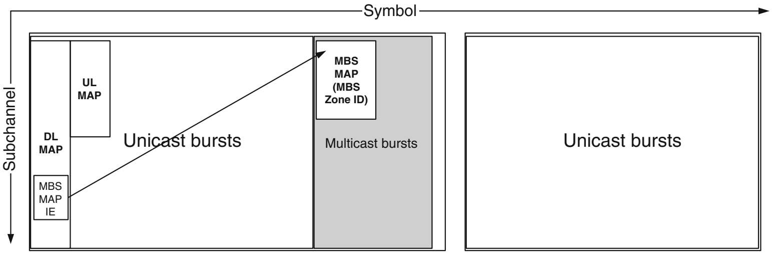


Fig. 8.17 MBS Frame Zone Layout

To support MBS, a special MBS zone is constructed within a frame. MBS capable base stations that belong to MBS zone have an identifier. Also, a MBS zone has its own MBS MAP to support more than one flow as seen in Fig. 8.17. All base stations belonging to the MBS Zone must be synchronized within an MBS Zone, to support multi-BS transmission.

The same way, macro-diversity transmission is also performed by setting the permutation formula to the same by fixing the cell_ID to 0 for all base stations who participates in macro diversity transmission. Macro diversity transmission implements a soft handover scheme in which a mobile station receives data transmission from multiple BS in either the same or different data region. MS performs diversity combining and soft combining.

8.3.6 Sounding Zone

In TDD, we mention that the uplink and downlink channels are reciprocal. But, if a BS is to derive the DL channel response from measured UL channel responses, the subscriber must first transmit on the UL, and that transmission must occupy the same portion of the bandwidth that will be used for the DL in the subsequent frame. This might be difficult to achieve in scenarios where the UL data traffic levels are significantly less than the DL traffic levels or when permutation zones are not reciprocal.

Sounding zones are designed to support the transmission of sounding signals by a subscriber to enable the BS to compute the UL channel response over the bandwidth. The sounding waveforms are designed to facilitate accurate UL channel estimation by the BS. The sounding instructions are communicated via a specific message transmitted in the UL-MAP. Notice that the sounding symbol may be independent of the following downlink data region allocation as seen in Fig. 8.18.

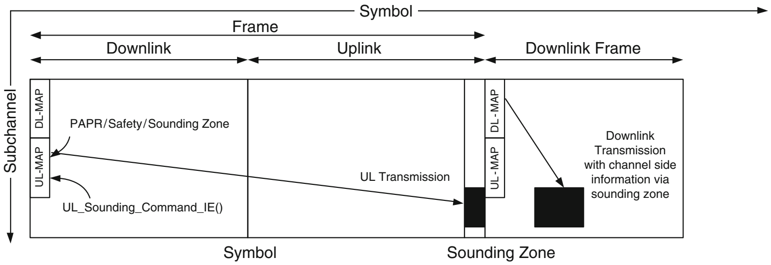


Fig. 8.18 Sounding Zone

8.4 Multiple Antenna System Support

IEEE 802.16e-2005 provides strong foundation for multiple antenna transmission and reception. Supported modes includes beamforming, transmit diversity, spatial multiplexing, and collaborative MIMO. Also, closed-loop operation is supported with various feedback mechanisms. This section discusses the use of multiple antenna techniques in mobile WiMAX. We first talk about the antenna array support. Then, we discuss MIMO systems that can be adapted for diversity, spatial multiplexing, and interference reduction.

8.4.1 Adaptive Antenna System

In WiMAX systems, several mechanisms are provided for Adaptive Antenna System (AAS). AAS mode in IEEE 802.16e-2005 can occur in AAS zones in DL and UL of the frame. AAS requires synchronous frame boundaries for all BS within the network and synchronous UL and DL ratio.

There are several attributes that are used in AAS operation. In UL, there is AAS preamble for training information and UL SDMA pilots. In DL, there are dedicated plots where MS performs the channel estimation. DL also provides AMC SINR measurement and SDMA pilots.

Training information is necessary for downlink. Either UL AAS preamble is used if the DL and UL allocation is symmetric, otherwise sounding zones can be used to train DL. The DL allocations must be preceded by UL training since additional latency can cause change of channel.

For stations that are at the cell edge, there is AAS DL zone that transmits the base station parameters in a robust transmission for initial ranging as well as paging and access allocation. An allocation is preceded with an AAS DL preamble. Although different beams can be transmitted within the zone, an allocation and associated preamble must be transmitted with the same beam.

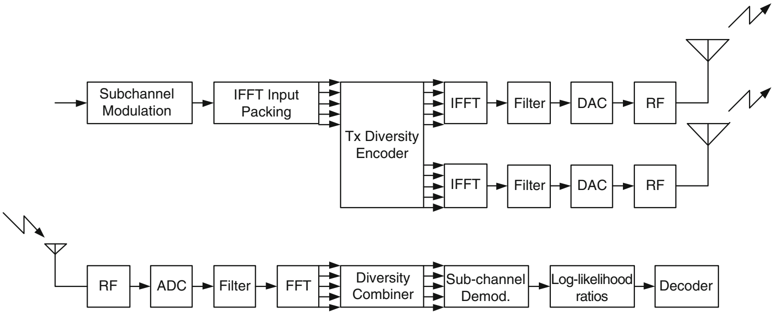


Fig. 8.19 Space-Time Coding

8.4.2 Space-Time Coding: Open-Loop

Space-Time Coding (STC) or frequency hopping diversity coding is used in the downlink to provide second order transmit diversity.

Figure 8.19 shows a transmitter with two transmit antennas and a receiver with one receive antenna. The PUSC mode of operation allows splitting the subchannels into segments. The transmit diversity mode of operation shall be used in a combined way where the regular subchannel and preamble transmission in the downlink shall be performed from only one antenna (Antenna 0) while the transmit diversity subchannel transmission shall be performed from both antennas.

Alamouti scheme is a basic scheme with two antennas. In the first channel use, antenna 0 transmits S_1, antenna 1 transmits S_2. In the second channel use, antenna 0 transmits $-S_2^*$, antenna 1 transmits S_1^* as in Matrix **A** represented equation below.

$$\begin{array}{cc} \mathbf{A} & \mathbf{B} \\ \begin{bmatrix} S_1 & -S_2^* \\ S_2 & S_1^* \end{bmatrix} & \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}, \\ Rate = 1 & Rate = 2 \end{array} \tag{8.3}$$

where Matrix **B** provides spatial multiplexing gain of 2 with no diversity.

Receiver gets r_0 (first channel use) and r_1 (second channel use) then estimates S_1 and S_2 with maximum likelihood estimator:

$$\begin{aligned} \hat{S}_1 &= h_0^* \cdot r_0 + h_1 \cdot r_1^* \\ \hat{S}_1 &= h_0^* \cdot r_0 + h_1 \cdot r_1^* \, . \end{aligned} \tag{8.4}$$

The STC transmission can be used both in PUSC and FUSC configurations. In PUSC, cluster structure is changed to fit the STC requirements. Two pilots are sent in the even symbols and pilots are shared by antenna 0 and 1 as seen in Fig. 8.20. In FUSC, the pilots within the symbols shall be divided between antennas. Antenna 0 uses variable set 0 and constant set 0 for even symbols and antenna 1 uses variable set 1 and constant set 1 for even symbols. They alternate in the odd symbols. Receiver waits for two symbols and combines them with maximum ratio combining.

STC can be extended to four antennas with following matrices;

$$\begin{array}{ccc} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \begin{bmatrix} S_1 & -S_2^* & 0 & 0 \\ S_2 & S_1^* & 0 & 0 \\ 0 & 0 & S_3 & -S_3^* \\ 0 & 0 & S_4 & S_3^* \end{bmatrix} & \begin{bmatrix} S_1 & -S_2^* & S_5 & -S_7^* \\ S_2 & S_1^* & S_6 & -S_8^* \\ S_3 & -S_4^* & S_7 & S_5^* \\ S_4 & S_3^* & S_8 & S_6^* \end{bmatrix} & \begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}, \\ Rate = 1 & Rate = 2 & Rate = 4 \end{array} \tag{8.5}$$

where Matrix **A** is to fulfill transmission diversity and Matrix **C** is for spatial multiplexing with rate 4; Matrix **B** is a hybrid scheme. Pilot structure for DL PUSC is depicted in Fig. 8.20 for four antennas. In FUSC, antennas time share the variable and constant sets as follows:

- In first symbol, Antenna 0 and 1 are active and use variable and constant set 0 and 1, respectively.
- In second symbol, Antenna 2 and 3 are active and use variable and constant set 0 and 1, respectively.
- In third symbol, Antenna 0 and 1 are active and use variable and constant set 1 and 0, respectively.
- In fourth symbol, Antenna 2 and 3 are active and use variable and constant set 1 and 0, respectively.

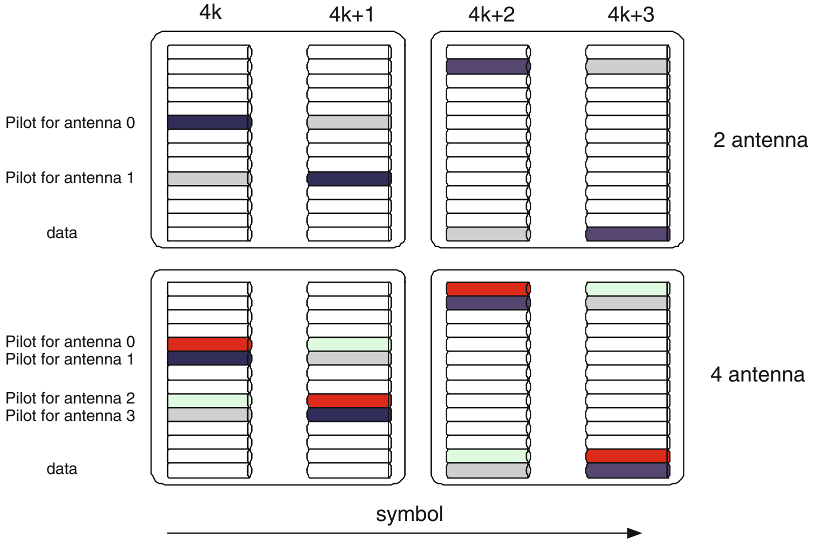


Fig. 8.20 Cluster structure for DL PUSC for 2 and 4 antennas

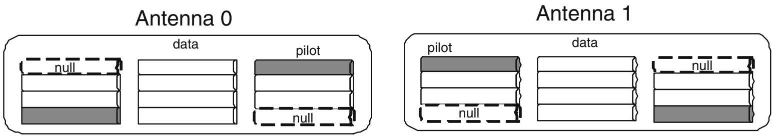


Fig. 8.21 Tile structure for UL PUSC for 2 antennas

Table 8.7 STC subpacket combining for 2-transmit antenna case

Initial	Odd	Even
$\begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$	$\begin{bmatrix} -S_2* \\ S_1* \end{bmatrix}$	$\begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$

Uplink STC also modifies the tile structure as in Fig. 8.21. Space-Time Transmit Diversity (STTD) or spatial multiplexing (SM) modes are applicable including collaborative spatial multiplexing. Supported modes for uplink are

- 2 antenna STTD
- 2 antenna SM with vertical encoding
- Single-antenna cooperative SM

Data subcarriers are encoded in pairs and for spatial multiplexing, vertical or horizontal mapping is applied in which either a single burst is modulated and then mapped or two bursts are modulated and individually mapped to antennas, respectively. For cooperative spatial multiplexing, a station uses only one pattern and other station complements the transmission by using the second pattern.

STC transmission also introduces a retransmission mechanism in two prong ways: STC incremental redundancy or chase combining. The mobile station combines the initial transmission with second retransmission in the form of space-time decoding. For incremental redundancy version, the combining for 2 transmit antenna case is shown in Table 8.7.

8.4.3 FHDC: Frequency Hopping Diversity Code

IEEE 802.16e-2005 defines optional diversity mode, known as Frequency-Hopping Diversity Code (FHDC) in which coding is done in space/frequency domain with two antennas. FHDC transmission is depicted in Fig. 8.22. In FHDC, first antenna transmits as regular transmission and second antenna encodes it with Alamouti scheme.

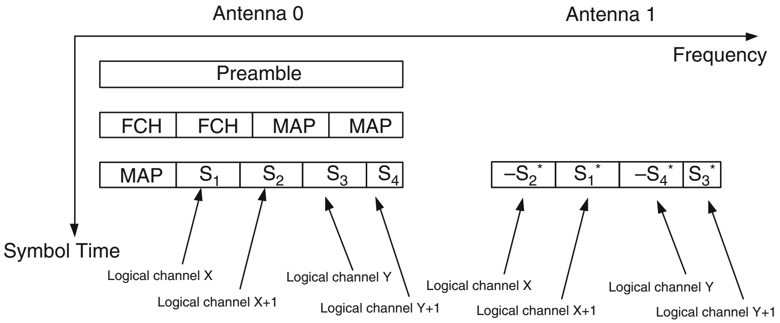


Fig. 8.22 Using FHDC in PUSC

Table 8.8 DL MIMO operation modes; vertical (V) encoding indicates transmitting one FEC stream over multiple antennas and horizontal (H) encoding indicates multiple FEC streams over multiple antennas

# antennas	Matrix	# layers	# station	encoding	rate	mapping
2	A	1	1	STDD	1	str0 to S1, S2
2	B	1	1	V	2	str0 to S1, S2
2	B	2	1	H	2	str0 to S1 str1 to S2
2	B	2	2	2	2	str0 to S1 str1 to S2
4	A	1	1	STDD	1	str0 to S1, S2
4	B	1	1	V	2	str0 to S1, S2
4	B	2	1	H	2	str0 to S1, S2
4	B	2	2	H	2	str0 to S1, S2
4	C	1	1	V	4	str0 to S1, S2
4	C	4	1	H	2	str0 to S1, S2
4	C	4	>1	H	4	str0 to S1, S2

8.4.4 MIMO: Closed-Loop

Besides STTD, MIMO operations include vertical and horizontal coding up to four antennas. Supported modes, listed in Table 8.8, require closed-loop system for better performance since closed-loop MIMO is designed to exploit CSI in order to tune the network to optimum with changing channel knowledge. Instantaneous and long-term channel statistics are used to determine the STC and precoding matrices of Fig. 8.23.

The CSI can be known in the transmitter either implicitly with channel reciprocity or explicitly with a feedback. The open-loop space-time coding is enriched with a MIMO precoding matrix in order to optimize the transmission. The space-time coding output ($\mathbf{x}$) is weighted by a matrix ($\mathbf{z} = \mathbf{W}\mathbf{x}$) before sending to transmit antennas as seen in Fig. 8.23.

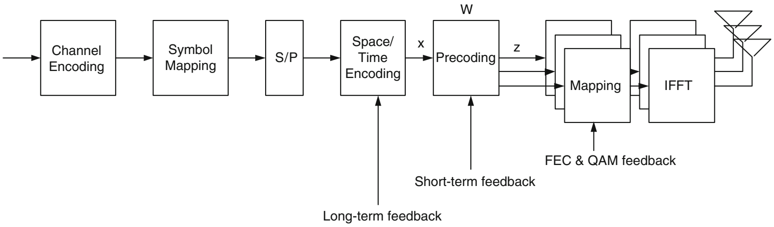


Fig. 8.23 MIMO precoding: note that **W** has $N_t \times s$ as dimension and **x** has $M_t \times 1$ as dimension, where N_t stands for the actual transmit antennas and s stands for number of streams at the output of the space-time coding

Two types of feedback is defined: short-term feedback and long-term feedback which are based on instantaneous channel knowledge and channel statistics respectively. Precoding scheme switches between a precoding matrix which works well with long-term channel properties and a precoding matrix which works well with short-term channel knowledge. Typically, short-term precoder is suitable in low mobility conditions and long-term precoder is used in high mobility. Whether station is capable or not for the feedback is negotiated in the network entry and mobile station informs base station about whether any of the following capabilities exist:

- Capable of calculating precoding weight
- Capable of adaptive rate control
- Capable of calculating channel matrix
- Capable of antenna grouping
- Capable of antenna selection
- Capable of codebook based precoding
- Capable of long-term precoding
- Capable of MIMO Midamble

8.4.4.1 Short-Term Closed-Loop Precoding

Short-term channel knowledge is one of the primary mechanism to determine the precoding matrix and requires heavy feedback. The matrix is selected from a codebook with 64 entries irrespective of the rank of the precoding matrix where each rank has 64 entry codebook. Short-term channel knowledge is used to fine tune the precoding matrix for changing channel conditions with short duration. If station is mobile then it may require to change its STC matrix as well since stations close to base station might utilize spatial multiplexing however stations away from base station might utilize transmit diversity. These changes are handled by long-term feedback.

8.4.4.2 Long-Term Closed-Loop Precoding

The long-term channel knowledge uses limited information such as channel mean (Ricean component) and covariance and is less accurate than instantaneous channel information. This feedback determines the spatial directions with respect to channel statistics. The number of spatial directions equals the rank of the precoder matrix, which determines the set from which the STC matrices are being chosen (Matrix A, B, or C). If the rank of the codebook is 3, then the STC matrices is chosen from the set of STC matrices for 4 transmit antennas. Let us clarify further, assume there are 4 transmit antennas and rank is 3. This indicates that matrices in the codebook has 4 rows and 3 columns. Depending on the spatial rate, one of the matrices is selected. For instance, for spatial rate 2, Matrix B is selected, which is designed to distribute the two spatial data streams to the three columns of the precoding matrix.

Long-term precoding also specifies the life-span of short-term precoding information and rank of codebook for long-term precoding matrix with its 6bit-CQICH feedback per user per second. Beyond the life span of the short-term precoding matrix, long-term precoding matrix is used until the next short-term precoding matrix is available to be used.

8.4.5 Feedback Methods

Fast-feedback channels are allocated to an MS to facilitate its response for closed-loop MIMO communication. Primary fast-feedback channel is an OFDMA slot in PUSC in which MS uses 48 data subcarriers to transmit 6-bits. The secondary fast-feedback channel is dedicated to transmit 4 bits.

MS feedbacks long-term codebook rank and long-term precoding matrix index as well as life span of short-term precoding matrix. BS typically selects the short-term precoding when desired and enforces long-term precoding if available otherwise.[6]

The available closed-loop MIMO feedbacks are listed below on the allocated channel quality indicator channel (CQICH) from subscriber station:

- **Antenna selection:** Mobile station uses secondary fast-feedback channel to indicate the transmit antennas that maximizes the channel capacity. The feedback includes number of streams (2bits), antennas selection index (3bits), and average CQI (5bits) of the selected antennas. For instance, for Matrix C for 4 transmit antennas, if number of streams is one, then one antenna out of four is selected with power boosting equals to one. For two streams, there are six choices in which two antennas are selected out of four with power boosting $1/\sqrt{2}$. Finally, for three streams, there are three choices in which three antennas out of four is selected with power boosting $1/\sqrt{3}$.
- **Antenna grouping:** Mobile station uses primary fast-feedback channel to indicate the number of transmit antennas to be used and the logical order of the antennas.The diversity matrices introduced before can be altered to change the

[6] Research shows that short-term precoding gains around 6–7 dB for fading rates below 10 Hz. However, long-term precoding gains from 1 to 5.5 dB depending on the antenna correlation and Rician K-factor.

transmission order. The feedback includes antenna grouping index (4-bits) and average CQI (5bits). For 4 antennas, Matrix A with Antenna grouping is depicted below:

$$\begin{array}{c} A_1 \\ \begin{bmatrix} S_1 & -S_2^* & 0 & 0 \\ S_2 & S_1^* & 0 & 0 \\ 0 & 0 & S_3 & -S_3^* \\ 0 & 0 & S_4 & S_3^* \end{bmatrix} \\ A_2 \\ \begin{bmatrix} S_1 & -S_2^* & 0 & 0 \\ 0 & 0 & S_3 & -S_4^* \\ S_2 & S_1^* & 0 & 0 \\ 0 & 0 & S_4 & S_3^* \end{bmatrix} \\ A_3 \\ \begin{bmatrix} S_1 & -S_2^* & 0 & 0 \\ 0 & 0 & S_3 & -S_4^* \\ 0 & 0 & S_4 & S_3^* \\ S_2 & S_1^* & 0 & 0 \end{bmatrix} \end{array} \tag{8.6}$$

- **Reduced Codebook based feedback:** Mobile station indicates the precoding matrix with either 3-bit or 6-bit feedback for the codebook with 8 entries or 64 entries, respectively. Also, includes average CQI (5bits) via secondary fast-feedback channel. Codebooks are defined for MIMO transmit beamforming to employ as the beamforming matrix in precoding stage. The notation $V(N_t, s, L)$ denotes the vector codebook where N_t denotes the dimension, s denotes the number of streams, and L is the number of bits required to indicate the feedback. Table 8.9 shows $V(2,2,3)$ codebook for 2×2 MIMO operation.

Table 8.9 3-bit codebook V(2,2,3)

Index	Column 1	Column 2
000	1 0	0 1
001	0.7940 −0.5801 + j0.1818	−0.5801 − j01818 −0.7940
010	0.7940 0.0576 + j0.6051	0.0576-j0.6051 −0.7940
011	0.7941 −0.2978 − j0.5298	−0.2978 + j0.5298 −0.7941
100	0.7941 0.6038 + j0.0689	0.6038 − *j*0.0689 −0.7941
101	0.3289 0.6614 + j0.6740	0.6614 − j0.6740 −0.3289
110	0.5112 0.4754 + j0.7160	0.4754 + j0.7160 −0.5112
111	0.3289 −0.8779 − j0.3481	−0.8779 + j0.3481 −0.3289

- **Quantized precoding weight feedback:** Mobile station quantizes the real and imaginary components of the MIMO channel and sends this information to BS with 6-bit feedback channel.

8.5 Channel Coding

Channel coding where components are shown in Fig. 8.24 comprises randomization, FEC encoding, HARQ if used, bit interleaving, and modulation. Optionally, repetition is performed before modulation. The binary data after randomization is fed into the encoder. Encoded data is bitwise interleaved and fed into the modulator which maps the bits to either QPSK, 16QAM, or 64QAM symbols. These symbols are assigned to subcarriers according to subcarrier allocation schemes described above sections. Subsequently, the OFDMA signal in time domain is computed via IFFT and cylic prefix is added afterwards.

8.5.1 Randomization

Randomizer aims to provide encryption over the air to avoid sniffing. Randomization is applied to each FEC block except preamble and FCH. Each data bit enters sequentially into the randomizer, MSB first. For HARQ, in each attempt, randomizer needs to be initialized to support joint decoding.

8.5.2 FEC Encoding

WirelessMAN OFDMA physical layer of IEEE 802.16-2004 together with IEEE 802.16e amendment supports several forward error correction (FEC) encoding. OFDMA standard has convolutional coding (CC) as the mandatory coding and specifies optional block turbo coding (BTC) and convolutional turbo coding (CTC). IEEE 802.16e amendment also has added low-density parity check coding (LDPC) as another optional coding scheme. IEEE 802.16e-2005 supports chase combining for all FEC types and incremental redundancy (IR) HARQ with CC and CTC.

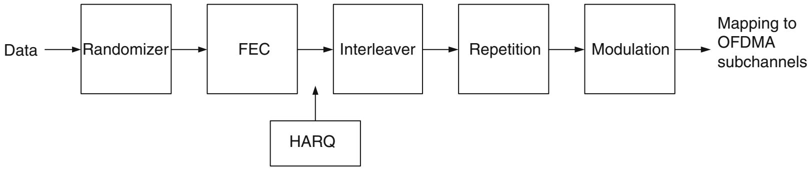


Fig. 8.24 Channel coding

In a frame, several HARQ channels are defined. Each channel may support several bursts, termed sub-bursts, which could have different HARQ modes. The available modes are listed below:

- HARQ-CC
- HARQ-IR for CRC
- HARQ-IR for CC
- MIMO HARQ-CC
- MIMO HARQ-IR for CC
- MIMO HARQ-STC

Multiple HARQ mode is supported and if there is a change, it is signaled in DL MAP. Additionally, zone boosting is supported if some subchannels are restricted to use in the DL PUSC mode. Consequently, used subcarriers are boosted according to the ratio of the number of useful subcarriers to others.

8.5.2.1 Convolutional Coding

Nonrecursive convolutional coding is the mandatory coding scheme in IEEE 802.16e-2005 with a rate 1/2. Constraint length is equal to $K = 7$ as seen in Fig. 8.25. Higher rate codes such as 2/3, 3/4, 5/6 are achieved with puncturing where the puncturing patterns are depicted in Table 8.10. In the receiver, signals are depunctured according to the specified pattern. Codec state is initialized with tail-biting where 6 bits from the last part of the sequence of the data block is appended to the front to flush out the remaining bits in the encoder. After encoding, 12 bits is discarded to remove the residue from the previous encoder.

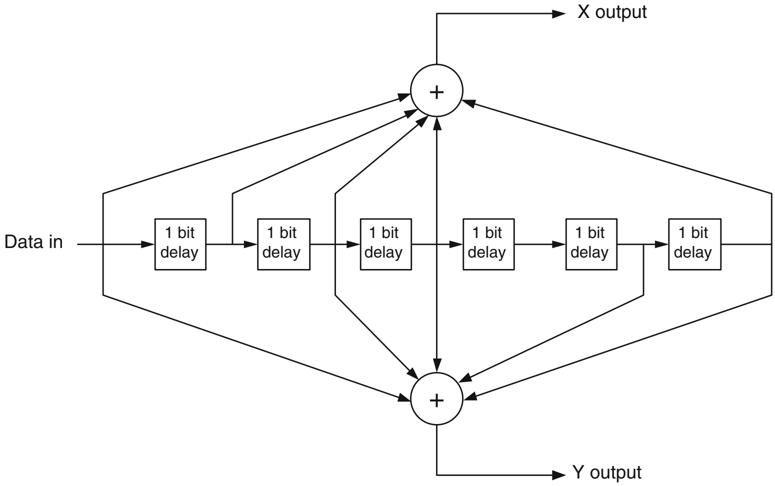


Fig. 8.25 Basic Convolutional Coding: generator polynomials are $G_1 = 171_{\text{OCT}}$ for X and $G_2 = 133_{\text{OCT}}$ for Y

Table 8.10 Puncturing patterns for convolutional coding

Code Rate	d_{free}	Parity 1 (X)	Parity 2 (Y)	Output
1/2	10	11	11	X_1Y_1
2/3	6	10	11	$X_1Y_1Y_2$
3/4	5	101	110	$X_1Y_1Y_2X_3$
5/6	4	10101	11010	$X_1Y_1Y_2X_3Y_4X_5$

Table 8.11 Puncturing patterns for HARQ-IR convolutional coding

SPID	0	0	1	1	2	2	3	3
	X	Y	X	Y	X	Y	X	Y
R 1/2	1	1	1	1	1	1	1	1
R 2/3	10	11	01	11	10	11	01	11
R 3/4	101	110	011	101	110	011	101	110
R 5/6	10101	11010	01011	10101	10110	01011	01101	10110

In HARQ-IR for CC, puncture pattern is defined to create HARQ packets with specified Suppacket ID (SPID). These patterns seen in Table 8.11 are used in the decoder to do the combination. The puncture pattern is the same for SPID $= 0$, and the rest is derived with cyclic shift of the previous one.

8.5.2.2 Zero-Tailed Convolutional Coding

Zero-tailed convolutional coding is another optional technique that enforces the encoder to return to the all-zero state by feeding a sufficient number of zeros at the end of each burst after randomization. This is rate deficient as compared to tail-biting convolutional codes.

8.5.2.3 Block Turbo Coding

Block turbo coding is one of the optional coding scheme defined in the IEEE 802.16e-2005 standard. Block turbo coding is product of two simple component codes such as extended Hamming codes or parity check codes. Coding polynomials are listed in Table 8.12.

First, row is encoded with (n_x, k_x) and then column is encoded with (n_y, k_y) as seen in Fig. 8.26. Thus, overall block size is $n_x \times n_y$ and information bits are $k_x \times k_y$ with the code rate of $R = R_x \times R_y$, where $R_i = k_i/n_i$. Transmission of the block over the channel occurs starting from the first row and followed by the second row, etc.

BTC may be shortened to fit the packet size as illustrated in Fig. 8.26. Hence, the rate becomes

$$R = \frac{(k_x - I_x)(k_y - I_y) - B - Q}{(n_x - I_x)(n_y - I_y) - B}, \tag{8.7}$$

Table 8.12 Hamming code generator polynomials

n'	k'	polynomial
15	11	X^4+X^1+1
31	26	X^5+X^2+1
63	57	X^6+X^1+1

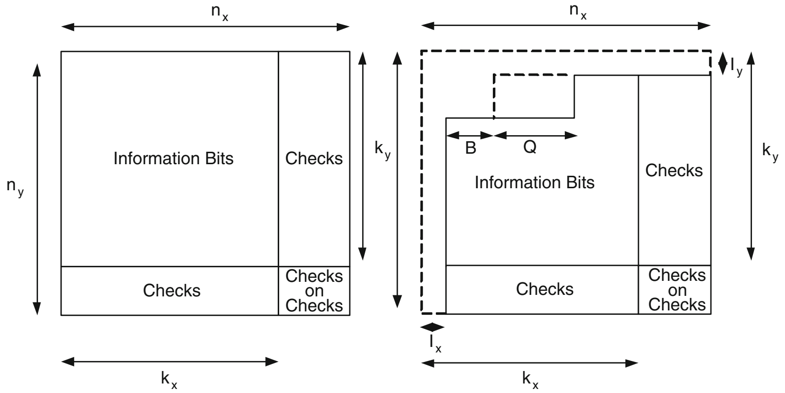


Fig. 8.26 BTC and shortened BTC structure

where I_x, I_y are removed columns and rows, respectively. B is removed individual bits and Q is leftover which is zero-filled by the encoder.

8.5.2.4 Convolutional Turbo Coding

Convolutional turbo coding is another optional coding scheme, which is high likely to be adopted by the WiMAX ecosystem. Figure 8.27 shows the CTC architecture, which uses double binary Circular Recursive Systematic Convolutional Code with a natural rate of $R = 1/3$. Input produces two encoding with and without interleaving. Later, by puncturing the mother code, subpackets, which are used for HARQ transmission, are generated with various coding rates. The constituent encoder has a natural rate of 2/4 and encoding procedure is as follows:

- Information bits (A and B) are fed to the output.
- First step, A and B are fed into the constituent encoder to produce parity bits Y_1 and W_1.
- Second step, A and B are fed into the constituent encoder after interleaving to produce parity bits Y_2 and W_2.
- Output code becomes $ABY_1W_1Y_2W_2$ when compared with input AB and the rate is 1/3.

Encoded symbols are demultiplexed into six subblocks as in Fig. 8.28. After that, subblocks are interleaved in two stages; first bits are flipped in alternating symbols;

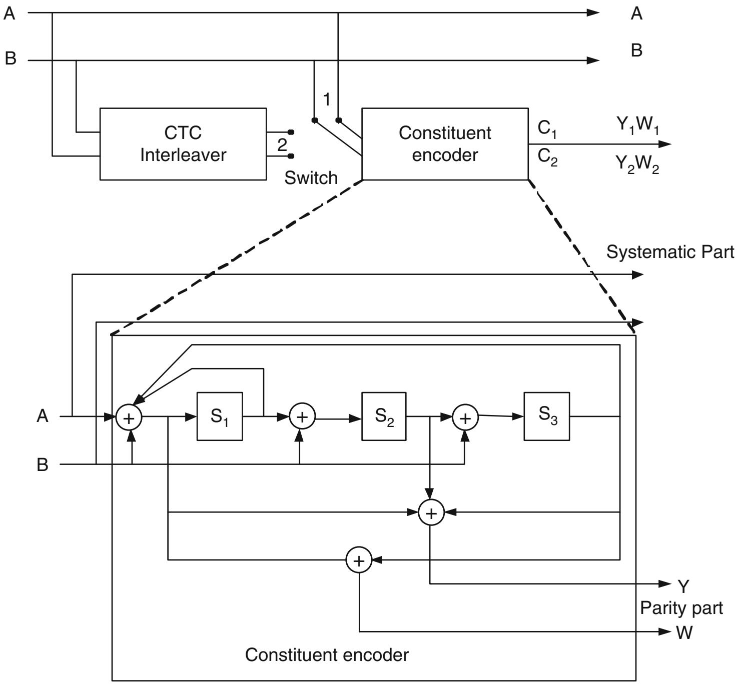


Fig. 8.27 Convolutional Turbo Coding

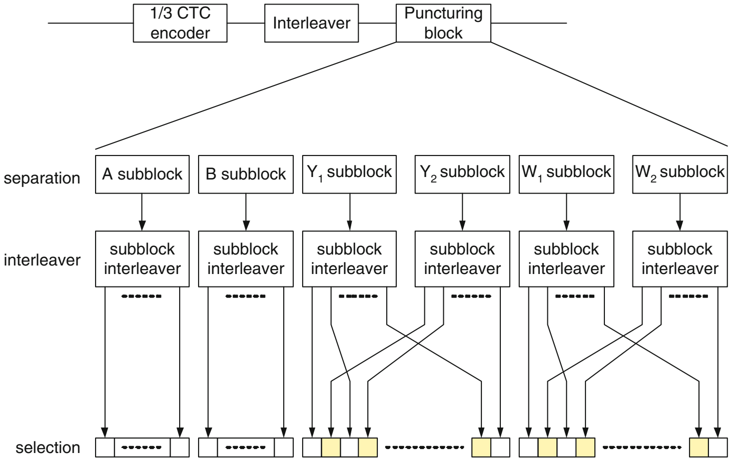


Fig. 8.28 Subpacket generation

then symbols are permuted; later, desired code rate is achieved with puncturing. Decoding uses tailbiting MAP algorithm to perform a-posteriori decoding.

Constituent encoder also uses tailbiting but now the encoder is recursive, which makes it not easy when compared with nonrecursive encoders. Initial encoding is introduced to ensure the starting state is the same as the ending state. The initial encoding starts in all-zero state and it ends up in a special state S. A look-up table is provided in IEEE 802.16 to relate the final state S and circulation state with regard to information sequence length. Circulation state is found given S produced by all-zero state. Finally, encoder is initialized with this circulation state.

In optional HARQ support, order is different than the case depicted in Fig. 8.24; when the size of packet is not in the allowed set of HARQ then '1's are padded. CRC-16 encoding is performed and included at the end of the padded and randomized packet. Randomization unlike previous case is performed on each allocation burst. Fragmentation occurs if size exceeds 4,800 bits and concatenation is applied if less. CTC encoding and subpacket generation is applied as explained above before modulation where mother code is transmitted with one of subpackets.

8.5.2.5 LDPC Coding

The LDPC coding is optional coding scheme, which is based on a set of systematic linear block codes where k information bits are encoded to n code bits by adding $m = n - k$ parity check bits. Each LDPC code is defined by a $\mathbf{H}$ matrix of size m-by-n, where m is the parity check bits in the code. $\mathbf{H}$ is composed of $\mathbf{P}_{i,j}$ matrices, which are z-by-z either permutation matrices or zero matrix, where $n = z.n_b$ & $m = z.m_b$ and $i \in \{0, m_b - 1\}$ & $j \in \{0, n_b - 1\}$.

$$\mathbf{H} = \begin{bmatrix} \mathbf{P}_{0,0} & \cdots & \mathbf{P}_{0,n_b-1} \\ \vdots & \ddots & \vdots \\ \mathbf{P}_{m_b-1,0} & \cdots & \mathbf{P}_{m_b-1,n_b-1} \end{bmatrix}. \tag{8.8}$$

As a result, base matrix $\mathbf{H}_b$ is composed of 1s and 0s where 1s are replaced by permutation matrices $\mathbf{P}_{i,j}$ and 0s are replaced by zero matrix to construct $\mathbf{H}$.

$\mathbf{H}_b$ is composed of systematic bits and parity-check bits. This flexibility supports different block sizes. Each base model matrix has $n_b = 24$ columns, and the expansion factor (z factor) is equal to $n/24$ for code length n where $z \in \{24, 28, 32, \cdots, 96\}$ and information bits are code rate R times n for $R \in \{1/2, 2/3, 3/4, 5/6\}$. Minimum code length ($n_{\min}$) is $24 \times 24 = 576$ bits and the maximum code length ($n_{\max}$) is $24 \times 96 = 2{,}304$ bits. There are five different base matrices, one for rate 1/2 coding, two for rate 2/3 coding and two for rate 3/4 coding.

8.5.3 Interleaving

All encoded data bits are interleaved by a block interleaver. First, interleaver maps the adjacent coded bits to nonadjacent subcarriers. Second, interleaver insures that adjacent coded bits are mapped alternately onto less or more significant bits of the constellation.

8.5.4 Repetition

Optionally, repetition coding is used to increase the robustness. Repetition factor (r) for 2,4, or 6 can be applied and allocated slots are provided to be the multiple of repetition factor for uplink and for downlink. Allocated slots are needed to be in the range of $\{r \times K, r \times K + (r-1)\}$, where K is the number of required slots before applying the repetition scheme. After FEC and interleaving, the data are mapped onto slots and each bit is repeated r times in contiguous slots. But, of course, due to randomization the data after constellation might differ.

8.5.5 Modulation

In the final stage, the **data modulation** is performed by entering bits serially to the constellation mapper. Gray-mapped QPSK and 16QAM is supported whereas 64QAM is optional. Adaptive modulation is used in the downlink and different modulation schemes for each subscriber is allowed in the uplink which are determined by the *MAC burst configuration messages* from the BS.

The constellation-mapped data are subsequently modulated onto the allocated subcarriers and each subcarrier multiplied by the factor $2(1/2 - w_k)$ according to the subcarrier index k and a sequence w_k is generated by the PRBS generator $(X^{11} + X^9 + 1)$. Also, this sequence is used for pilot, preamble/midamble and ranging modulation:

- **Pilot modulation** is performed with $Re(c_\mathrm{k}) = 2(1/2 - w_k)$ and $Im(c_k) = 0$ in the uplink where c_k is the kth subcarrier of the ranging channel. In the downlink and for the optional uplink tile structure, each pilot is transmitted according to $Re(c_k) = 8/3(1/2 - w_k)$ and $Im(c_k) = 0$ with a boosting of 2.5 dB over the average power of each data tone.
- **Preambles/midambles** are modulated according to $Re(preamble) = 4 \cdot \sqrt{2} \cdot (1/2 - w_k)$ and $Im(preamble) = 0$.
- **Ranging** modulation is defined by $Re(c_k) = 2 \cdot (1/2 - C_k)$ and $Im(c_k) = 0$ where C_k is the k^{th} bit of the code generated.

8.6 Control Mechanisms

Physical layer also introduces certain mechanisms to control the air link transmission. Also, these control features are ingredients to perform end-to-end management. Control mechanisms defined in the standard are

- Ranging
- Power control
- Channel Quality Measurements

Before talking about them in detail, as a side note, we need to talk about synchronization. Since BSs are required to be time synchronized for TDD and FDD duplexing. The synchronizing reference is typically provided by GPS with a 1 pps timing pulse and a 10 MHz frequency reference or by network timing protocols such as IEEE 1588, etc. MS also acquires and adjusts its timing to have BS receive all uplink OFDMA symbols time coincident.

Although MS and BS operates with the same clock in a time synchronized network, MS might be in various distance to the BS. Additional time and power measurements are used to alleviate near-far problem and propagation delay. Ranging is a procedure defined in the standard to perform time and power adjustments. Ranging solely is not sufficient to determine the power since channel quality measurement is needed to understand whether the allocated power is ample for a transmission to reach to the BS. Combination of all these also provides necessary information for higher layer protocols to execute network entry, mobility, and paging features.

8.6.1 Ranging

Ranging is a collection of processes by which the MS and BS maintain the quality of the RF communication link between them. The process is based on MS transmitting a binary-coded signal and BS responding with required adjustments such as frequency, time, and power.

Ranging codes are BPSK modulated Pseudo-Noise (PN) sequences, produced by the PrBS.[7] The length of the sequence is 144 bits and there are 256 orthogonal codes. A nonoverlapping set of ranging codes are allocated to initial ranging, periodic ranging, bandwidth requests, and handover ranging. This information is broadcasted by the base station in each cell and adjacent cells use nonoverlapping sets to prevent collision of codes.

MAC layer defines a ranging channel that is composed of one or more groups of six (eight) adjacent subchannels for UL PUSC (UL optional PUSC and AMC). MS finds the location of the ranging channel from UL-MAP and transmits a ranging code from a bank of codes to perform ranging. Note that more than one MS may use the ranging channel simultaneously but codes are resolved in the BS as long as they are orthogonal.

8.6.1.1 Initial/Handover Ranging

Initial ranging allows an MS, joining the network, to acquire correct transmission parameters, such as time offset and transmitter power level, so that MS can communicate with the BS.

[7] Pseudorandom Binary Sequence is initialized by UL_IDcell and constructed by the polynomial generator $1+X^1+X^4+X^7+X^{15}$.

The initial ranging is performed during two symbols where same ranging code is transmitted in each symbol. Initial ranging can also be performed over four symbols in which two different codes are transmitted over two symbols. Also, ranging code over a symbol is possible, which is typically constructed by applying IFFT to BPSK modulated binary ranging code and appending the guard interval.

8.6.1.2 Uplink Periodic Ranging and Bandwidth Request

Because of rapid changing channel, MS needs to periodically adjust uplink transmission. Therefore, MS performs periodic ranging to update time, frequency, and power.

Also, MS may request a bandwidth from BS by performing bandwidth request. Bandwidth request transmission is performed by sending one code over one symbol or three code over three consecutive symbols. For each uplink bandwidth grant, BS seeks for a transmitted signal. If there is no signal BS terminates otherwise depending on the quality of the signal, either sends "continue" or "abort" message in ranging response.

MS processes each ranging response and implements the correction as indicated. If ranging response is success, the MS shall use the grant to service its pending uplink data queues. If MS has not been given the opportunity to transmit after the expiration of a timer maintained by MS. MS assumes that the link is broken.

The downlink burst profile is determined by the BS according to the quality of the signal that is received from each MS. If the received CINR goes outside of the allowed operating region, the MS requests a change to a new burst profile.

8.6.2 Power Control

The BS provides accurate power measurements of the received burst signal to feed back to mobile station as a calibration message. Mobile station maintains the same power density where total transmitted power changes proportionally with the number of active subchannels assigned. Mobile subscriber also informs the base station about the maximum available power and normalized transmitted power to have base station select the optimum coding and modulation schemes.

There are two types of power control: open-loop and closed-loop power control. Open loop power control is the procedure where the mobile station (aka SS) adjusts its transmit power according to received signal level without explicit instruction by the BS. The station set power value according to

$$P(\mathrm{dBm}) = L + C/N + NI - 10\log_{10}(R) + \mathrm{offset}_{\mathrm{SS}} + \mathrm{offset}_{\mathrm{BS}}, \tag{8.9}$$

where P is the transmit power level per subcarrier, L is the estimated average current uplink propagation loss and calculated based on the total power received on the active subcarriers of the preamble, C/N is the normalized C/N of the modulation/FEC rate for the current transmission, R is the number of repetitions for the modulation/FEC rate, NI is the estimated average power level (dBm) of the noise and interference per subcarrier at BS, offset$_{\mathrm{SS}}$ is the correction term for SS-specific power offset and controlled by SS, offset$_{\mathrm{BS}}$ is the correction term for BS-specific power offset and controlled by BS.

There are two types of open loop power control: passive and active where passive control sets offset$_{\mathrm{SS}}$ to zero and active control adjusts offset$_{\mathrm{SS}}$ within a range.

Open loop procedure is fast since it works without round-trip delay between the base station and the user terminals. The main disadvantage is the limited correlation between received power level on the uplink and downlink.

Closed-loop power control defines a feedback mechanism for BS to measure the received signal and make adjustment procedures by sending to the mobile station to adjust output power. The delay between measurement and application is critical. Closed-loop power control utilizes periodic ranging and bandwidth request for adjustment. Once MS fails to receive ranging response after sending a periodic ranging code, MS may adjust its transmit power for the subsequent periodic ranging codes. Similarly, for bandwidth request ranging, once MS fails to receive a bandwidth allocation after sending a bandwidth request code, MS may adjust its transmit power for the subsequent bandwidth request ranging codes.

8.6.3 Channel Quality Measurements

There are two metrics for channel quality measurements: RSSI and CINR. MS reports the mean and the standard deviation of the RSSI and CINR in units of dBm. One possible method to estimate the RSSI is given by

$$\mathrm{RSSI} = 10^{-\frac{G_{rf}}{10}} \frac{1.2567 \times 10^4 V_{\mathrm{c}}^2}{(2^{2B})R} \left(\frac{1}{N} \sum_{n=0}^{N-1} |Y_{IorQ}[k,n]| \right)^2 mW, \tag{8.10}$$

where B is Analog-to-Digital Convertor (ADC) precision, number of bits of ADC; R is ADC input resistance [Ohm]; V_{c} is ADC input clip level [Volts]; G_{rt} is analog gain from antenna connector to ADC input; $Y_{IorQ}[k,n]$ is nth sample at the ADC output of I or Q branch within signal k; N is number of samples.

To estimate the CINR, the ratio of the sum of signal power and the sum of residual error for each data sample is computed using the following equation

$$CINR[k] = \frac{\sum_{n=0}^{N-1} |s[k,n]|^2}{\sum_{n=0}^{N-1} |r[k,n] - s[k,n]|^2}, \tag{8.11}$$

where $r[k,n]$ received sample n within message k and $s[k,n]$ is the pilot sample.

8.7 Summary

We gave a detailed overview of OFDMA PHY of IEEE 802.16e-2005, which is the selected physical layer of mobile WiMAX. OFDMA PHY brings great level of flexibility with various possible configuration options:

- Adjacent and distributed subcarrier permutation formulas are defined to formulate subcarrier to subchannel mapping for various diversity options. FUSC, PUSC, TUSC, AMC are possible configurations.
- Each subcarrier permutation defines a slot structure, which is the basic building block of an OFDMA frame.
- Currently, a TDD mode is defined in IEEE 802.16e and FDD mode is being designed together with WiMAX Forum.
- A frame has a MAP for downlink and uplink to indicate the burst allocation of a user.
- Various coding schemes with or without HARQ are supported including turbo coding and LDPC coding.
- OFDMA PHY supports extensively multi antenna operation: AAS, MIMO systems can be configured for diversity, beamforming, and spatial multiplexing.

References

1. *IEEE Standard 802.16-2004, Part 16: Air interface for fixed broadband wireless access systems,* June 2004.
2. *IEEE Standard 802.16e-2005, Part 16: Air interface for fixed and mobile broadband wireless access systems,* December 2005.
3. Burr, A., "Turbo-codes: the ultimate error control codes?" *Electronics and Communication Engineering Journal,* pp. 155–165, 2001.
4. Bahl, L. R., Cocke, J., Jelinek, F., Raviv, J., "Optimal Decoding of Linear Codes for Minimizing Symbol Error Rate," *IEEE Trans. on Information Theory,* IT-20, pp. 284–287, 1974.
5. Berrou, C., Glavieux, A., Thitimajshima, P., "Near Shannon limit error-correcting coding and decoding," *Proceedings of the ICC,* pp. 1064–1070, 1993.
6. Lindskog, E., Artes, H., Tujkovic, D., Rath, K., Shashidhar, V., Rajan, B. S., Vaze, R., Bergkvist, A., Lorenz, B., Mandava, B., Paulraj, A., Agrawal, A., "Closed Loop MIMO Precoding," IEEE C802.16d-04/293r2, 2004.

Chapter 9
WiMAX MAC Layer

The IEEE 802.16 standard is designed to provide a system for fixed and mobile broadband wireless access. Mobility operation is added with IEEE 802.16e-2005 amendment for frequencies below 6 GHz.

In the previous chapter, we discussed the OFDMA physical layer of the standard. Now, we introduce the medium access layer (MAC) where it defines access in two ways: point-to-multipoint (PMP) and mesh type of communication. PMP access is a one-hop wireless communication between a base station (BS) that has a direct connection to the network and plurality of mobile stations (MSs). Downlink transmission from BS to MSs is broadcast and uplink transmissions from MSs to BS are unicast. BS acts as the central point to facilitate downlink and uplink access for a particular mobile station. BS allocates burst regions within a frame for an MS to send or receive data. MS knows these allocated regions by decoding the MAP portion of each frame where it can learn the burst region that is destined to it in the downlink, and to find out the dedicated region in the uplink for it to place a transmission.

In IEEE 802.16e-2005, access is connection oriented and each flow is associated by a transport connection. For each type of transport connection, there are three functions: service flow creation, maintenance, and termination. During creation, certain quality of service (QoS) parameters are associated with the connection. Initial bandwidth allocation may be altered due to stimulus from either MS or the network with maintenance function. Finally, connections may be terminated when a customer's service requirement changes.

Mesh-type access, on the other hand, enables a MS to service other MSs. As a result, there could be MSs that are more than one hop away from the BS. The stations that are one node away are called *neighbors*. Neighbors of a station are called *neighborhood*. An extended neighborhood contains all the neighbors of the neighborhood. Unlike PMP BS, mesh BS and other stations in the two-hop neighborhood coordinate for access decisions to avoid collision in this said two-hop neighborhood. PMP BS is capable of handling multiple independent sectors with sectorized antenna. On the other hand, mesh systems are deployed with omnidirectional antennas.

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_9, © Springer Science+Business Media LLC 2009

In this chapter, we focus more on PMP mode, which is the adopted MAC mode of WiMAX, and briefly discuss mesh mode in the last chapter under IEEE 802.16j section. First, we introduce the MAC reference model and introduce the packet formats. Second, we talk about QoS, network entry, handover, and paging modes. Finally, we conclude the chapter with key highlights.

9.1 Reference Model

To understand the MAC layer, it is better to introduce the WiMAX reference layer in this chapter. Figure 9.1 indicates the key network components that are mentioned in this chapter. More details about WiMAX architecture will be given in the next chapter.

As you can see from the figure, a base station has two interfaces: air interface and network. Air interface requires PHY and MAC modules and network interface requires a network module. Therefore, MAC is responsible to mediate the communication between PHY and network module. There are three sublayers that assist MAC in this process as depicted in Fig. 9.2. The service-specific convergence sublayer (CS) acts as an intermediate hop to transform the external network data, received from CS service access point (SAP), into MAC service data units (SDUs). These are received by the MAC common part sublayer (CPS) through the MAC SAP.

MAC CPS is flexible enough to interface with multiple CS specifications. Consequently, this layered architecture isolates the need of any specific information related to internal of CS. CS is mainly responsible to associate the SDU to the proper MAC service flow identifier (SFID[1]) and connection identifier (CID[2]). IEEE 802.16e-2005 provides following CSs:

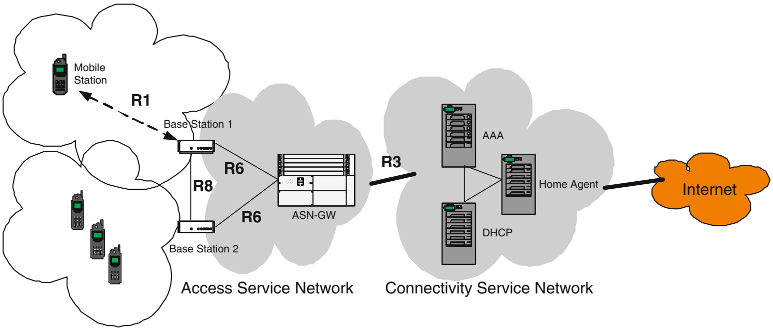


Fig. 9.1 WiMAX reference model: R1 is a reference point for WiMAX PHY; R6 is a reference point between BS and ASN Gateway; R8 is a reference point between BSs; R3 is a reference point between ASN Gateway and connectivity service network

[1] Remains constant within ASN while session is active.

[2] Changes with the serving BS.

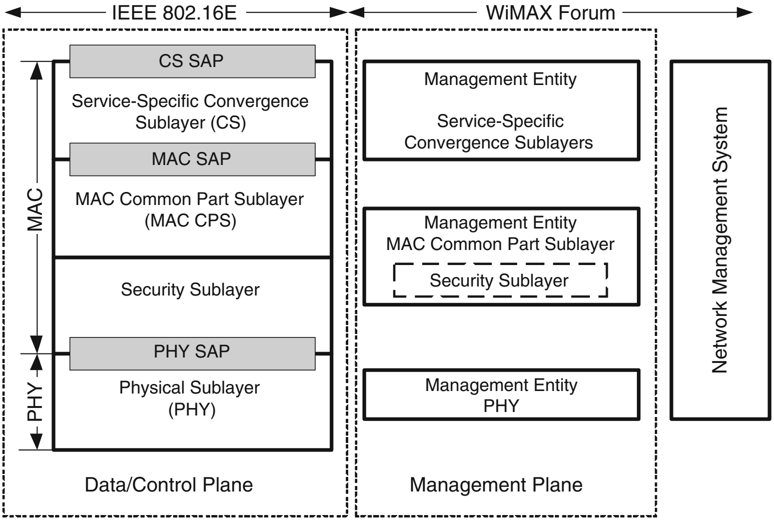

IEEE 802.16E
WiMAX Forum
CS SAP
Service-Specific Convergence Sublayer (CS)
MAC SAP
MAC Common Part Sublayer (MAC CPS)
Security Sublayer
PHY SAP
Physical Sublayer (PHY)
MAC
PHY
Management Entity Service-Specific Convergence Sublayers
Management Entity MAC Common Part Sublayer
Security Sublayer
Management Entity PHY
Network Management System
Data/Control Plane
Management Plane

Fig. 9.2 Protocol layering

0. ATM CS
1. Packet CS IPv4
2. Packet CS IPv6
3. Packet CS Ethernet (802.3)
4. Packet CS VLAN (802.1/Q)
5. Packet CS IPv4 over Ethernet
6. Packet CS IPv6 over Ethernet
7. Packet CS IPv4 over VLAN
8. Packet CS IPv6 over VLAN
9. Packet CS 802.3 with optional VLAN tags and ROHC header compression
10. Packet CS 802.3 with optional VLAN tags and ERTCP header compression
11. Packet IPv4 with ROHC header compression
12. Packet IPv6 with ROHC header compression

MAC CPS is the core of the MAC that accommodates multiple access, resource reservation, connection establishment, and maintenance functions. MAC also maintains the security architecture and PHY SAP feature that facilitate the transmission of the data from the MAC CPS to the air.

CS layer classifies higher layer data and associates with a specific MAC connection where there are rules associated in order to characterize the QoS of the service flow. Data are encapsulated into the MAC SDU format with optional 8-bit payload header suppression index (PHSI) (see next section). BS implements a scheduler that has a priority mechanism to order the transmission time of the individual SDUs, and classification is performed in BS for downlink and in MS for uplink packets.

9.2 PHS: Packet Header Suppression

Header suppression is used to suppress repetitive part of the header to decrease the overhead since certain fields remain unchanged during a session. For example, if IP session is established, source and destination IP addresses remain unchanged. These fields can be replaced by a value in the transmitter and then inserted back in the receiver. Header suppression increases the efficiency significantly when packet sizes are very small such as in VoIP.

PHS is the packet header suppression scheme defined in IEEE 802.16 standard but optional in WiMAX. We also introduce RObust Header Compression (ROHC) in the next chapter, which is considered by the WiMAX Forum as the header suppression scheme.

The PHS operation is set up during the dynamic service flow creation between BS and MS. There are several predefined PHS rules to be used in this operation. A PHS rule contains all parameters related to header suppression of the SDU. The PHS rule is not application agnostic and might change with respect to the type of the service, such as VoIP, HTTP, FTP, etc. When CS receives a SDU from higher layer, it maps the SDU into a SFID and CID and extracts the associated PHS rule, if any. From the PHS rule, fields to be suppressed are found in PHS field (PHSF) and fields not to be suppressed are found in PHS mask (PHSM). Suppressed MAC SDU is appended by a PHS index (PHSI) as provided by the PHS rule; also verification is possible with PHS verify (PHSV), which basically compares the PHSF with what is expected according to the PHS rule. If PHSF of the SDU does not match, then PHSI is set to 0. In the receiver side, PHSI is used to extract PHSF and PHSM from the PHS rule in order to reconstruct the suppressed portion.

9.3 Data/Control Plane

Now, we start describing the data and control plane modules. Both modules require identifiers to differentiate the connections. Typically, MS has a 48-bit MAC address that is unique globally and also BS assigns a 16-bit connection identifier (CID) per connection. When MS first enters the network, two pairs of management CIDs are provided by the BS via initial ranging procedure. Third pair is established optionally for *managed* MS, which allows network to control some of its parameters. A CID pair per connection comprises two unidirectional CIDs, each dedicated either for uplink or downlink. Three management CIDs are classified according to their functions as follows:

- *Basic management connection: Basic CID* is used when BS and MS exchange short, time-urgent MAC management messages.
- *Primary management connection: Primary CID* is used when BS and MS exchange longer, more delay-tolerant MAC management messages.

- *Secondary management connection: Secondary CID* is used to transfer delay tolerant, standards-based messages such as DHCP, TFTP, SNMP, etc.

Note that management CIDs are not used to transfer data. During session creation, a CID pair is assigned by BS per data bearer.

9.3.1 MAC PDU Formats

The CID is carried in the MAC packet data unit (PDU) header (see Fig. 9.3). MAC PDU is the basic payload unit that inherits the IEEE 802.16e specific features. SDUs arriving from the higher layer are mapped into PDU in the MAC CPS to be transmitted over the air. There are various types of PDU header for downlink and uplink.

9.3.1.1 MAC PDU Headers

Downlink MAC only uses one type of header, which is called *generic MAC header* (GMH) as seen in Fig. 9.4. GMH is the first portion of a MAC PDU that carries either management messages or SDUs. For uplink, there are two types of headers: The first one is GMH to carry either management messages or CS data. The other

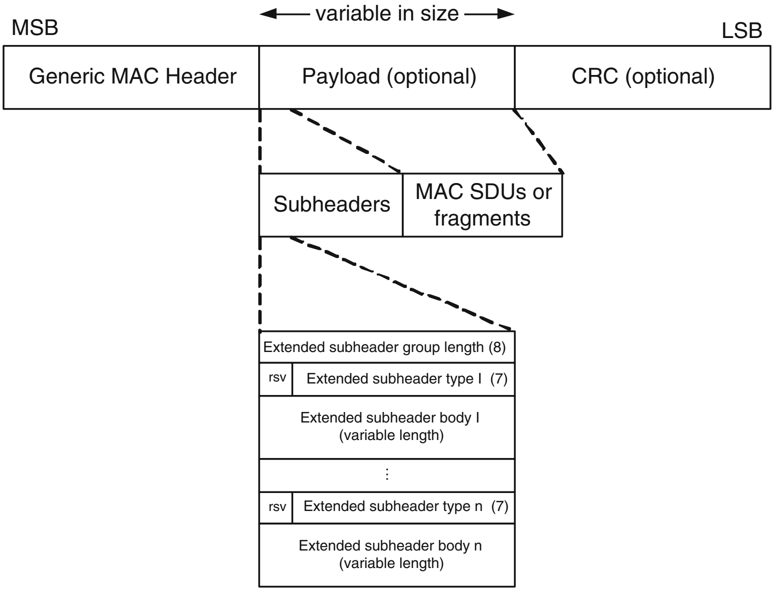

variable in size
MSB
LSB
Generic MAC Header
Payload (optional)
CRC (optional)
Subheaders
MAC SDUs or fragments
Extended subheader group length (8)
rsv
Extended subheader type I (7)
Extended subheader body I (variable length)
⋮
rsv
Extended subheader type n (7)
Extended subheader body n (variable length)

Fig. 9.3 MAC PDU formats: PDU and CRC are optional. After generic MAC header, there could be subheaders and after subheaders there could be MAC SDUs or fragments thereof. Subheader types are ordered, and extended subheader precedes the rest

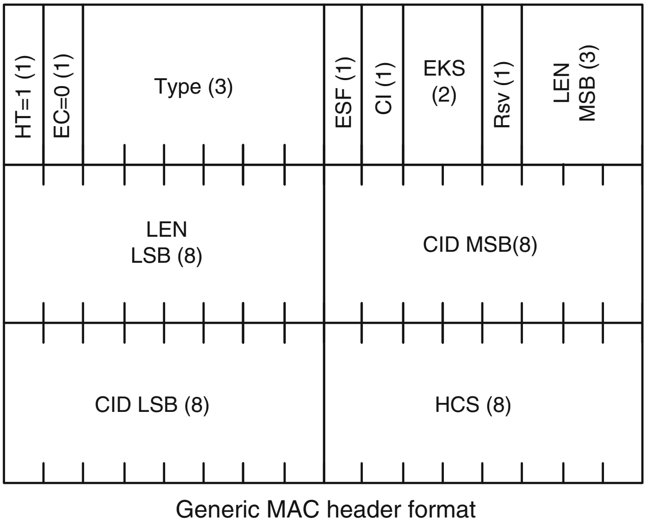


Fig. 9.4 Generic MAC header format: header type (HT) and encryption control (EC) together determine the type field that indicates the subheaders and special payload types. Extended subheader field (ESF) indicates the presence of extended subheader after the header. Extended subheaders are not encrypted. CI is CRC Indicator; CID is connection identifier; EKS is encryption key sequence, which is the index of the traffic encryption key (TEK) and initialization vector to be used to encrypt the payload. HCS is header check sequence and LEN is length in bytes of the MAC PDU

type is seen in Fig. 9.5 and called *MAC header without payload*, which does not carry any information. *MAC header without payload* serves several purposes:

Type 1: *Bandwidth request header* is used for uplink communication and includes the CID to indicate the connection for which uplink bandwidth is requested and the number of bytes requested.

Type 1: *Bandwidth request and UL Tx power report header* includes additionally the uplink transmit power in dBm for the burst that carries this header.

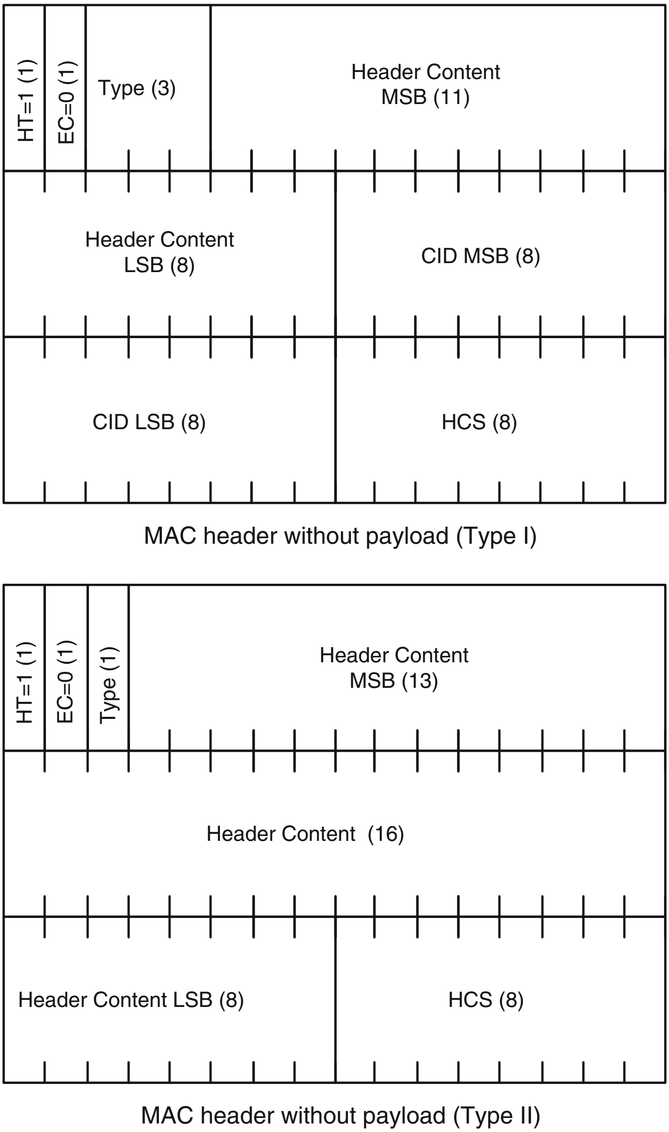
Type 1: *Bandwidth request and CINR report header* includes CINR measurement of MS in addition to the bandwidth request.

Type 1: *CQICH allocation request header* is allocated by the BS so that MS reports continuously CINR of BS preamble. As a result, BS can tune the appropriate modulation and coding.

Type 1: *PHY channel report header* is used by the MS to report the uplink transmit power level in dBm for the burst that carries this header.

Type 1: *Bandwidth request and uplink sleep control header* also requests de/ activation of certain power saving class in addition to incremental transmission demand.

Type 1: *SN report header* is used in handover to provide the last sequence number for each flow. Three sequence numbers can be presented in this header and this can be repeated one more time; consequently, up to six flows can be supported. Sequence number can be obtained from ARQ or virtual MAC SDU sequence number.



MAC header without payload (Type I)

MAC header without payload (Type II)

Fig. 9.5 MAC header without payload formats

Type 2: *Feedback header* is sent by an MS either as a response to *feedback polling IE* or the *feedback request extended subheader* or as an unsolicited feedback. Content can be DL average CINR, UL transmission power, MIMO information, etc.

Type 2: *MIMO channel feedback header* is used for MS to provide DL MIMO channel quality feedback to the BS. Content can be type or number of antennas, precoding matrix, CQI, etc.

A CRC can be added depending on the service flow requirement to cover the GMH and the payload. If there is encryption, the CRC is calculated after encryption. CRC-32 is used for OFDMA mode with the standard generator polynomial of degree 32:

$$\begin{aligned} G(x) = {} & x^{32} + x^{26} + x^{23} + x^{22} + x^{16} \\ & + x^{12} + x^{11} + x^{10} + x^{8} + x^{7} \\ & + x^{5} + x^{4} + x^{2} + x + 1. \end{aligned} \tag{9.1}$$

Payload encryption is defined by a security association (SA) but GMH is not allowed to be encrypted since it contains the necessary information for decryption. To decrypt a payload, *EKS* field contains the sequence number of the periodically refreshed key. At each generation of keying material, BS increments the two-bit sequence number. MS keeps the last two keying materials for uninterrupted service during an SA's key transition.

9.3.1.2 MAC Subheaders

Also, there are subheaders that can succeed after the GMH.

- Fragmentation subheader
- Grant management subheader
- Packing subheader
- ARQ feedback subheader
- Mesh subheader
- Fast-feedback allocation subheader

These can be followed by one or multiple extended subheaders:

DL *SDU SN extended subheader* contains the last virtual MAC SDU sequence number of current MAC PDU.

DL *DL sleep control extended subheader* is sent by BS to activate/deactivate certain power saving class.

DL *Feedback request extended subheader* provides a UL allocation for a fast feedback channel transmission.

UL *MIMO mode feedback extended subheader* is used by MS to provide MIMO mode feedback.

UL *UL Tx power report extended subheader* is used to report the Tx power of the uplink burst that carries this subheader.

UL *Minifeedback extended subheader* is used in uplink to carry the feedback content.

DL *SN request extended subheader* is requested by BS from the MS in order to have it send the SN report header via this subheader.

DL/UL *PDU SN (short/long) extended subheader* is to specify the PDU sequence number in a monotonic increasing manner.

9.3.1.3 MAC Management Messages

MAC management messages are carried in the payload of the MAC PDU via basic, primary, and secondary CIDs or broadcast. Only management messages that are carried via primary CID can be packed and/or fragmented. MAC management messages serve several purposes. First type of MAC management messages are used to inform the MSs by the BS about cell configuration or frame configuration:

- UCD: Uplink channel descriptor is transmitted by the BS to inform the physical characteristic of the uplink channel. It informs if there is a change in the configuration or in the allocated regions for initial ranging and bandwidth request. [Fragmentable Broadcast]
- DCD: Downlink channel descriptor is transmitted by the BS at a periodic interval to define the characteristics of a downlink physical channel. DCD also contains downlink burst profile. [Fragmentable Broadcast]
- DL-MAP: Downlink MAP contains DL-MAP information elements (DL-MAP IEs), which basically address the portion of the frame for a particular downlink connection. [Broadcast]
- UL-MAP: Uplink MAP contains the uplink bandwidth allocations. UL-MAP message contains at least one IE that marks the end of the last allocated burst. [Broadcast]

The second type of MAC management messages are used to perform network entry:

- RNG-REQ/RSP: Ranging request and response are used at initialization and periodically to determine network delay and to request power and/or downlink burst profile change. [Initial Ranging or Basic]
- SBC-REQ/RSP: MS basic capabilities message is transmitted during initialization. MS sends its basic capabilities in request and BS responses with the supported ones. [Basic]
- PKM-REQ/RSP: Privacy key management request and response are used at the authentication. There are several formats including authentication including PKMv2, RSA, etc. [Primary]
- REG-REQ/RSP: Registration request and response are used at the registration to exchange MAC address, IP version, CS and MS capabilities, etc. [Primary]
- TFTP-CPLT/RSP: MS receives the configuration file from the provisioning server. [Primary]

The third type of messages are used to establish, maintain, and terminate service flows:

- DSA-REQ/RSP/ACK: Dynamic service addition is used to create new service flows. It can be initiated by network as well as MS. Service flow parameters are set and Service flow ID (SFID) is assigned with CID if the connection is active by network. [Primary]
- DSC-REQ/RSP/ACK: Dynamic service change is used by an MS or BS to dynamically change the parameters of an existing service flow. [Primary]

- DSD-REQ/RSP: Dynamic service deletion is used by an MS or BS to delete an existing service flow. [Primary]
- MCA-REQ/RSP: Multicast polling assignment messages are used to assign or remove an MS from a multicast polling group. [Primary]
- DBPC-REQ/RSP: Downlink burst profile change messages are initiated by MS to request a change of the downlink burst profile used by the BS to transport data to the MS. [Basic]
- DSX-RVD: DSx received message is sent by BS to MS to indicate that DSx-REQ sent by MS has been received in a more timely manner than provided by the DSx-RSP message. [Primary]

These messages are complemented with other type of messages that are used to maintain robust connectivity with the BS:

- ARQ-feedback/discard/reset: ARQ feedback message is used to signal any combination of different ARQ ACKs. ARQ Discard message is used to skip a certain number of ARQ blocks. ARQ Reset message is sent to reset the parent connection's ARQ transmitter and receiver state machines. [Basic]
- REP-REQ/RSP: Channel measurement report request is sent by BS to require RSSI and CINR channel measurement reports. Report response contains the report. [Basic]
- FPC: Fast power control message is used by BS to adjust the power levels of multiple subscribers simultaneously. [Broadcast]
- CLK-CMP: Clock comparison message is to synchronize clock signals of the network. [Broadcast]

Also, there are messages introduced for MIMO and AAS support:

- AAS-FBCK-REQ/RSP: Adaptive antenna system channel FeedBack messages serve to obtain channel measurement that will help in adjusting the direction of the adaptive array. [Basic]
- AAS_Beam_Select: AAS beam select is sent unsolicited by MS to inform BS about the preferred beam for the AAS MS sending this message. [Basic]
- AAS_BEAM_REQ/RSP: AAS beam messages are used by a system supporting AAS to request channel measurement that will help in adjusting the direction of the adaptive array. [Basic]
- PRC-LT-CTRL: The BS can establish a long-term precoding by sending this to the MS. Same message can be used to tear down the long-term precoding. [Basic]

Mobility support is provided with a set of messages to address scanning, handover (HO), paging, and idle mode:

- MOB_SLP-REQ/RSP: MS sends Mobile Sleep Request message to de/activate certain Power Saving Class. Response sent by BS may contain the definition of Power Saving Class or Power_Saving_Class_ID unique to a MS. [Basic]
- MOB_TRF-IND: Traffic Indication Message is sent from BS to MS to indicate whether there has been traffic addressed to each MS that is in sleep mode. An MS that is in sleep mode, during its listening-window decodes this message to find out an indication addressed to itself. [Broadcast]

- MOB_NBR_ADV: Neighbor Advertisement Message is transmitted periodically by BS to identify the network and define the characteristics of neighbor BS to potential MS seeking initial network entry or handover. [Broadcast, Primary]
- MOB_SCN-REQ/RSP: Scanning Interval Allocation Request may be transmitted by an MS to request a scanning interval for the purpose of seeking available BSs and determining their suitability as targets for HO. In response to request, BS transmits a response to initiate scan reporting. [Basic]
- MOB_BSHO-REQ/RSP: BS HO Request message is sent by BS to initiate HO. BS HO Response is sent upon receiving MOB_MSHO-REQ. [Basic]
- MOB_MSHO-REQ: MS HO Request message is transmitted by MS to initiate HO. [Basic]
- MOB_HO-IND: HO Indication is sent by MS to indicate the final decision that it is about to perform an HO. MS may cancel or reject as well. [Basic]
- MOB_SCN-REP: Scanning report is to report the scanning results to its serving BS after each scanning period at the time indicated in the MOB_SCN-RSP message. [Primary]
- MOB_PAG-ADV: BS Broadcast Paging message is sent with Broadcast CID during the BS Paging Interval. [Broadcast]
- MBS_MAP: BS may send a Multicast Broadcast Service message to describe the Multicast Broadcast Connections. MBS_MAP contains MBS_MAP_IEs. [–]
- PMC_REQ/RSP: Power control Mode Change request and response messages are to change the power control mode between the open-loop power control and closed-loop power control. [Basic]
- MOB_ASC-REP: Association Result Report is used for association level 2 in which scanning report is gathered in the backbone and sent back to MS from its servicing BS. [Primary]

Finally, there is a subset of messages that allow control of MS via BS:

- RES-CMD: Reset Command is transmitted by BS to force the MS to reset itself and reinitialize its MAC. This is used if there is continued uplink abnormalities or when MS is unresponsive. [Basic]
- DREG-CMP: De/reregister Command message is used to send action codes to MS. Actions are change channel, listen but not transmit, listen but transmit only management signals, revoke normal operation, and terminate normal operation. [Basic]
- DREG-REQ: De/reregister Command is sent by BS to force the MS to change its state. MS can change the current channel, suspend transmission, only transmit management messages, resume normal operation, or terminate normal operations. [Basic]
- MSH-NCFG/NENT/DSCH/CSCH/CSCF: MeSH Network ConFiGuration message is a basic handshake between nodes to see their vendor, operator, etc. Network ENTry message is used for synchronization and initial network entry into a Mesh network. Distributed SCHedule messages shall be transmitted in Mesh mode when using distributed scheduling. Centralized SCHeduling message is used for centralized scheduling to request bandwidth from the Mesh BS.

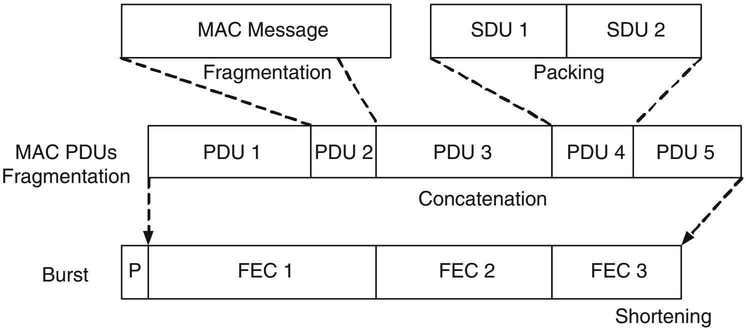


Fig. 9.6 Payload configuration

Centralized Scheduling ConFiguration is broadcast to all neighbors, and all nodes rebroadcast the message according to its index number specified in the message. [Broadcast/NENT is basic]

9.3.2 Construction and Transmission of MAC PDUs

Management messages and data packets are subject to three ways of packet manipulation: fragmentation, concatenation, and packing as seen in Fig. 9.6. Fragmentation divides a MAC SDU into one or more MAC PDUs. This allows great flexibility to increase the efficiency of the frame as well as QoS scheduling. If ARQ is enabled, fragmented packet is retransmitted. Otherwise, the fragments are transmitted once and in sequence. According to sequence number, the receiver detects a loss and discards all MAC PDUs until a PDU with first fragment bit set or an unfragmented MAC PDU. Other way around is also possible where MAC PDUs can be concatenated into a single transmission. Each MAC PDU is identified by a unique CID; therefore, received packet can be decomposed easily. Another flexibility in packet manipulation is packing where multiple MAC SDUs can be packed into a single PDU. If the SDU is of fixed size then the number of SDUs packed can be identified from the length field of MAC header. If SDUs are of variable size then a subheader is appended to each SDU. These features can be used simultaneously to utilize air-link more efficiently with certain guidelines. Information can be carried either in the packing subheader (PSH) or fragmentation subheader (FSH) when no PSH is present.

9.3.3 ARQ Mechanism

ARQ mechanism considered in the standard is per connection basis where a MAC SDU is partitioned into blocks whose length is specified by the ARQ_BLOCK_SIZE. Selected blocks for retransmission are encapsulated into a PDU.

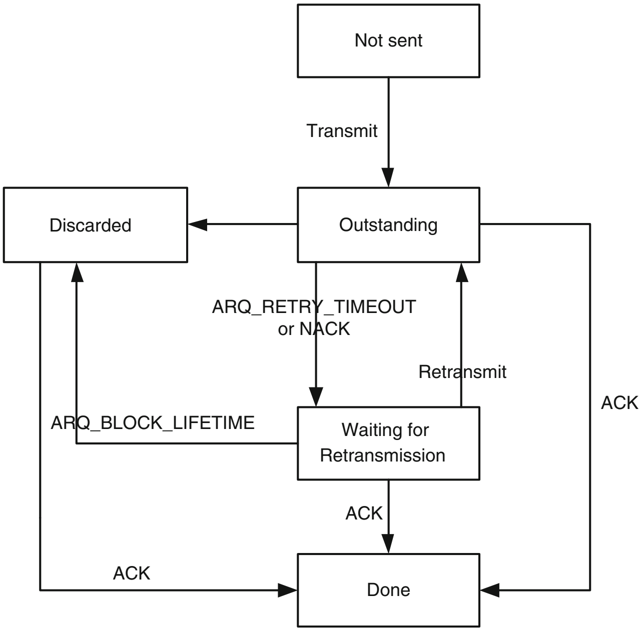


Fig. 9.7 Transmitter state machine

Each block has a Block Sequence Number (BSN), which is mapped to a normalized value by the following equation:

$$\text{bsn}' = (\text{bsn} - \text{BSN\_base}) \bmod \text{ARQ\_BSN\_MODULUS}. \qquad (9.2)$$

FSM for transmitter is illustrated in Fig. 9.7 where only blocks that are only in the ARQ_WINDOW_SIZE are transmitted as long as ACK is received. If ACK is not received for a certain block, that block is retransmitted until it exceeds the block lifetime; otherwise, it is discarded.

There is a flexibility to acknowledge all blocks with Cumulative ACK. If Cumulative ACK is used, when a valid BSN is received, the transmitter considers all blocks in between the ARQ_TX_WINDOW_START to BSN as acknowledged and set window start to BSN+1.

Also, there is Selective ACK that reports for individual blocks. If Selective ACK is received, transmitter considers the blocks as acknowledged according to entries in a bitmap. ARQ_TX_WINDOW_START is incremented when received BSN is equal to ARQ_TX_WINDOW_START. However, if Cumulative with Selective ACK is received, actions for Cumulative ACK are performed and then Selective ACK.

Receiver side is illustrated in Fig. 9.8 where only blocks within an interval (defined by ARQ_RX_WINDOW_START and ARQ_WINDOW_SIZE) are accepted. The ARQ_RX_WINDOW_START is incremented to the next lowest numbered ARQ block with block reception. When all ARQ blocks of the MAC SDU have been correctly received, MAC SDU is handled to upper layers.

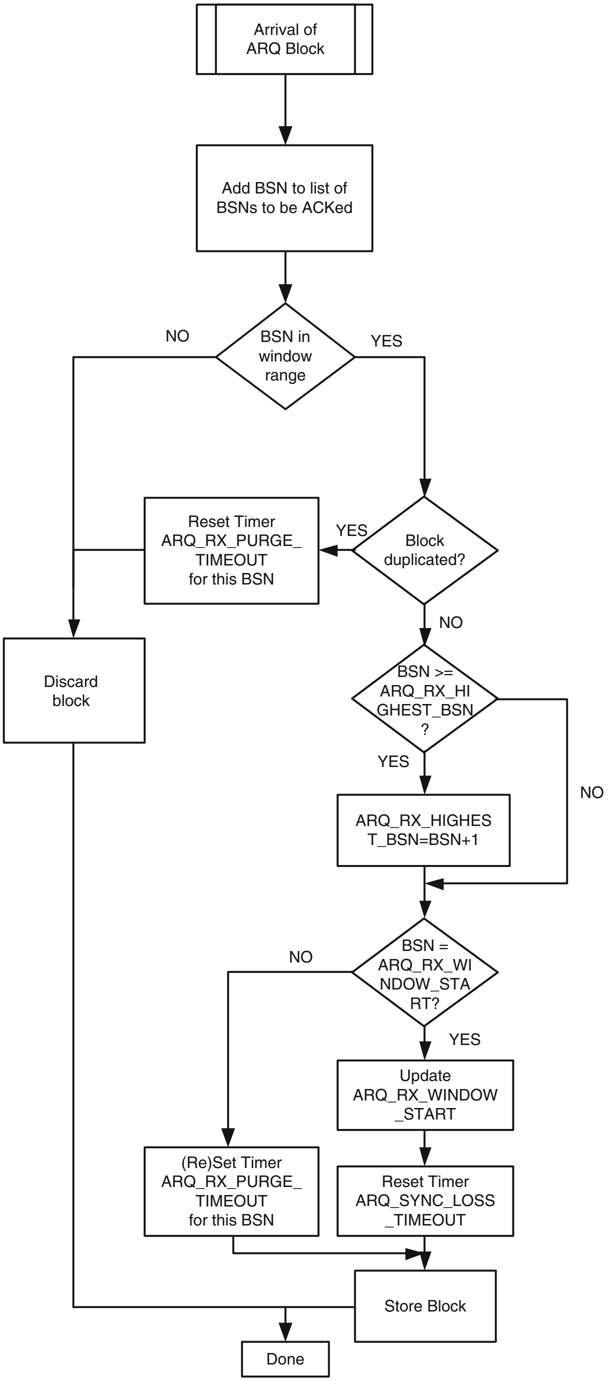

Arrival of ARQ Block
Add BSN to list of BSNs to be ACKed
BSN in window range
NO
YES
Reset Timer ARQ_RX_PURGE_ TIMEOUT for this BSN
YES
Block duplicated?
NO
Discard block
BSN >= ARQ_RX_HI GHEST_BSN ?
YES
NO
ARQ_RX_HIGHES T_BSN=BSN+1
BSN = ARQ_RX_WI NDOW_STA RT?
NO
YES
Update ARQ_RX_WINDOW _START
(Re)Set Timer ARQ_RX_PURGE_ TIMEOUT for this BSN
Reset Timer ARQ_SYNC_LOSS _TIMEOUT
Store Block
Done

Fig. 9.8 Receiver state machine

9.3.4 Transmission Scheduling

Having introduced the construction and transmission of PDUs, we now talk about transmission ordering, aka scheduling. Each data connection is associated with a set of QoS parameters, which are defined through DSA or DSC signaling. DL and UL service flow types differ but complement each other. Tables 9.1 and 9.2 illustrate the mandatory parameters for downlink and uplink service flow types, respectively. Defined service flows are as follows for uplink and downlink:

UGS: Unsolicited Grant Service is for real-time data streams with fixed-size data packets issued at periodic intervals (T1/E1 and VoIP without silence suppression, etc.). The grant is periodic and eliminates the need of a bandwidth request. This reduces latency but if grants are not used, the bandwidth is wasted. [DL and UL]

rtPS: Real-time Polling Service is for real-time data streams with variable-sized data packets that are issued at periodic intervals (MPEG video, etc.). Since the grants are for variable-sized packets, the BS provides opportunities for unicast polling. This increases overhead but it gives opportunity to pass on the grant if there are no data to transmit; otherwise, this provides a bandwidth to fit the size. This unicast polling opportunity also reduces the response latency to accommodate stringent delay requirements of the multimedia sessions. [UL]

Table 9.1 DL-supported QoS parameters

QoS parameter	UGS	RT-VR	NRT-VR	BE
Maximum sustained rate (bps)	M	M	M	M
Maximum reserved rate (bps)	O (=MSR)	M	M	ns
Maximum tolerable rate (bps)	ns	ns	ns	ns
Maximum jitter (ms)	M	ns	ns	ns
Priority	ns	M	M	M
Maximum latency (ms)	M	M	ns	ns
Maximum burst size (bytes)	0	0	0	0

Table 9.2 UL-supported QoS parameters

QoS parameter	UGS	rtPS	nrtPS	BE
Maximum sustained rate (bps)	M	M	M	M
Maximum reserved rate (bps)	O (=MSR)	M	M	ns
Maximum tolerable rate (bps)	ns	ns	ns	ns
Maximum jitter (ms)	M	ns	ns	ns
Priority	ns	ns	M	M
Maximum latency (ms)	M	M	ns	ns
Maximum burst size (bytes)	0	0	0	0
Unsolicited polling interval (ms)	M	M	ns	ns

RT-VR: Real-Time Variable Rate service is to support real-time data applications with variable bit rates. This service guarantees data rate and delay. [DL]

nrtPS: Nonreal-Time Polling Service is for delay-tolerant data streams consisting of variable-sized data packets for which a minimum data rate is required (FTP, etc.). In this service, MS can use either contention-based access to request for a grant or unicast polling opportunities similar to rtPS but with longer periods. [UL]

NRT-VR: Nonreal-Time Variable Rate Service supports applications that require a guaranteed data rate but not sensitive delay requirements. [DL]

ertPS: Extended Real-Time Polling Service is a hybrid scheme based on UGS and rtPS. This is designed in 802.16e to introduce flexibility to packet sizes of UGS by giving option to MS to transmit in the dedicated periodic grant either for a data transmission or an additional bandwidth request. This way a transmission size can be variable in size as in rtPS by using the periodic grants like an unicast polling opportunity. [UL]

ERT-VR: Extended Real-Time Variable Rate Service is for applications that require guaranteed data and delay such as VoIP with silence suppression. [DL]

BE: Best Effort is for data streams with no requirement on minimum service level. MS uses contention-based polling in uplink to request for a bandwidth. [DL and UL]

Transmission scheduling, performed by BS, assigns data to a particular bandwidth allocation in a frame. BS is aware of downlink and uplink bandwidth requests and grants bandwidth opportunity to a MS. Unlike downlink packets that arrive from network, uplink bandwidth is requested by MS via either a bandwidth request header or a piggyback request. These requests are made in terms of number of bytes needed to carry the MAC header and payload, but not the PHY overhead. Also, these requests are either incremental or aggregate. Incremental request is added to the previous bandwidth need but a new aggregate request replaces the current bandwidth request. Note that piggyback bandwidth requests are always incremental. Moreover, aggregate bandwidth request is sent periodically to do self-correction.

OFDMA PHY introduces also contention-based CDMA bandwidth request where MS picks a ranging code and transmits in the ranging subchannel. BS responds with CDMA_Allocation_IE to specify the transmit region and transmitted ranging code via broadcast. There is a possibility of collision by the selection of the same code and region by more than one MS. In that case, contention resolution protocol is applied. Each bandwidth request is addressed to the MS's Basic CID and MS may use it for any of its connections. If the granted bandwidth does not satisfy the need, it can rerequest or discard the SDU.

Also, polling-based bandwidth request is available where BS polls individual MS through unicast polling to give the opportunity to MS in the UL-MAP to send a bandwidth request. However, if the connection is UGS, MS is not polled unless PM (Poll-me-Bit) is set. If there is less bandwidth than the amount necessary for individual polling, BS may poll groups through multicast CID. In this case, only MS that needs bandwidth replies and if there is a collision, it is resolved through contention resolution protocol.

Contention resolution protocol is based on a truncated binary exponential backoff where initial and the maximum backoff window sizes are controlled by the BS through UCD messages. MS randomly selects a number within its backoff window and defers. During wait, if it receives a unicast *Request IE* or *Data Grant Burst Type IE*, MS stops the contention resolution and transmits again. If transmission is unsuccessful, MS increases the backoff interval and randomly selects a number within its new backoff window.

9.4 Network Entry and Initialization

Up to now, we discussed how packets are handled in MAC layer. From this section, we start introducing the control signaling. An MS is accepted to the WiMAX network after performing the network entry procedure. MS first scans for downlink channel and establishes synchronization with the BS of the selected network. Transmission parameters are acquired from the UCD message and ranging with the BS is performed accordingly. Next step is negotiating the basic capabilities that are hosted in the network. This is followed by authentication and registration. Registration is followed by IP address acquisition, which is the final step in network entry procedure. This makes MS ready to establish service flows either preprovisioned or dynamic. Figure 9.9 illustrates the flow chart for network entry and we describe it in detail in the following:

- *Network Discovery and Selection:* MS scans for the base stations and discovers the network access providers (NAPs) (advertised by SII-ADV). A NAP can serve for multiple network service providers (NSPs) – operators. MS has a manual or automated way of selecting the NAP and NSP.
- *Obtaining Downlink/Uplink Parameters:* BS generates UCD and DCD messages on the downlink at periodic intervals. BS also generates UL-MAP and DL-MAP at intervals as specified in a particular PHY specification, typically per frame. MS synchronizes with a NAP and extracts the parameters through these messages. After synchronization, MS waits for UCD message to retrieve a set of transmission parameters for uplink. MS waits for a bandwidth allocation map for the selected channel and begins its first uplink transmission.
- *Initial Ranging:* Ranging is the process of acquiring the correct timing offset and power adjustments such that MS's transmission is aligned to the received frame. For the OFDMA PHY, the initial ranging process shall begin by sending initial-ranging CDMA code on a randomly selected ranging slot, which is the UL allocation dedicated for that purpose. MS sends CDMA code and waits for RNG-RSP message. If MS does not receive a response, MS sends a new CDMA code at one-step higher power level. When BS receives a CDMA code, BS cannot tell which MS sent that ranging request therefore BS broadcasts a RNG-RSP

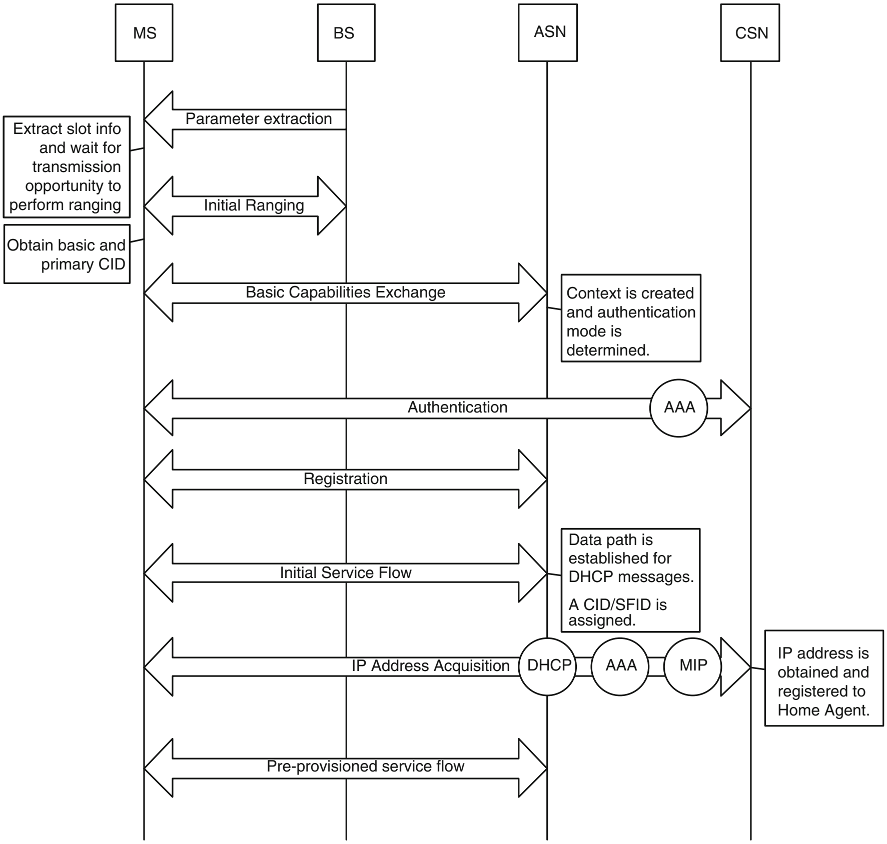

MS
BS
ASN
CSN
Parameter extraction
Extract slot info and wait for transmission opportunity to perform ranging
Initial Ranging
Obtain basic and primary CID
Basic Capabilities Exchange
Context is created and authentication mode is determined.
Authentication
AAA
Registration
Initial Service Flow
Data path is established for DHCP messages.
A CID/SFID is assigned.
IP Address Acquisition
DHCP
AAA
MIP
IP address is obtained and registered to Home Agent.
Pre-provisioned service flow

Fig. 9.9 Network entry diagram

that advertises the ranging slot (OFDMA symbol number, subchannel) where the CDMA ranging code has been identified and the received CDMA code in order to address specifically the MS. Then, BS provides a BS allocation for the MS using the CDMA_Allocation_IE in order to have it send a RNG-REQ message again. BS ends initial ranging by replying with RNG-RSP that contains the basic and primary CID to be assigned to the MS.

- *Basic Capabilities:* MS and the network need to negotiate the capabilities to find out the lowest common denominator. In SBC-REQ, MS advertises supported PHY parameters, supported bandwidth allocation parameters, security negotiation parameters, capabilities for construction and transmission of MAC PDUs, power save class types capability, supported association type, etc. BS receiving SBC-REQ from MS informs the network about the new MS entering the network. Network creates a new context block for the MS and responds with supported capabilities. These plus the BS's capabilities are sent to MS via SBC-RSP message.

- *Authentication:* Network triggers the authentication process after negotiation of basic capabilities. The EAP authentication process is performed between MS and Authentication Server via Authenticator in ASN-GW. MS and Authentication Server negotiate the EAP method and perform it. Authenticator is in pass-through mode and forwards the EAP message to AAA server using RADIUS messages. After this authentication process key holder in the ASN-GW receives the keys for data transport and mobility.
- *Registration:* Now, MS is ready to enter the network. MS sends REG-REQ to initiate registration. REG-REQ carries IP version, CS capabilities, mobility parameters, and information about handover support, etc. BS exchanges information with the network and responds with REG-RSP, which is formatted according to BS policy and/or network response.
- *Initial Service Flow:* Initial Service Flow (ISF) is required to establish the signaling for IP configuration. A CID/SFID is associated to process the uplink and downlink IP packets.
- *IP Address Acquisition:* There can be several ways to obtain the IP address. IP address can be provided by AAA, HA, or DHCP server. If HA is used then there needs to be a DHCP proxy in the ASN where a DHCP-Discover message is sent to. DHCP Proxy triggers the PMIP client to initiate Mobile IP procedure; PMIP client sends this message to the Foreign Agent (FA) in order to initiate the registration to Home Agent (HA). HA provides the IP address HoA and this is relayed back to the MS. If CMIP is used, CMIP residing in MS talks to FA directly. Also, DHCP proxy can get the IP address from AAA and register that to the HA. If DHCP server provides the IP address then a DHCP relay is needed in the ASN.
- *Preprovisioned Service Flow:* There can be preprovisioned service flows that need to be activated after MS registration. Authenticator receives the indication of successful completion of authentication from AAA server, and Service Flow Authorization (SFA) module detects the completion of registration. SFA starts the service flows according to QoS profiles of the MS. SFA sends the request to BS and BS relays this with DSA-Request signal where MS responds with DSA-RSP.

9.5 QoS

Network entry is followed by service flow establishment. A service flow operation is defined in the QoS framework, which introduces configuration, registration, and signaling functionalities in order to dynamically enable the service flows and adjust the traffic parameters.

The primary purpose of the QoS is to define ordering and scheduling on the air interface (R1) as we talked before. But, in order to provide end-to-end QoS, scheduling mechanism on the air interface should be complemented with mechanisms residing in the network.

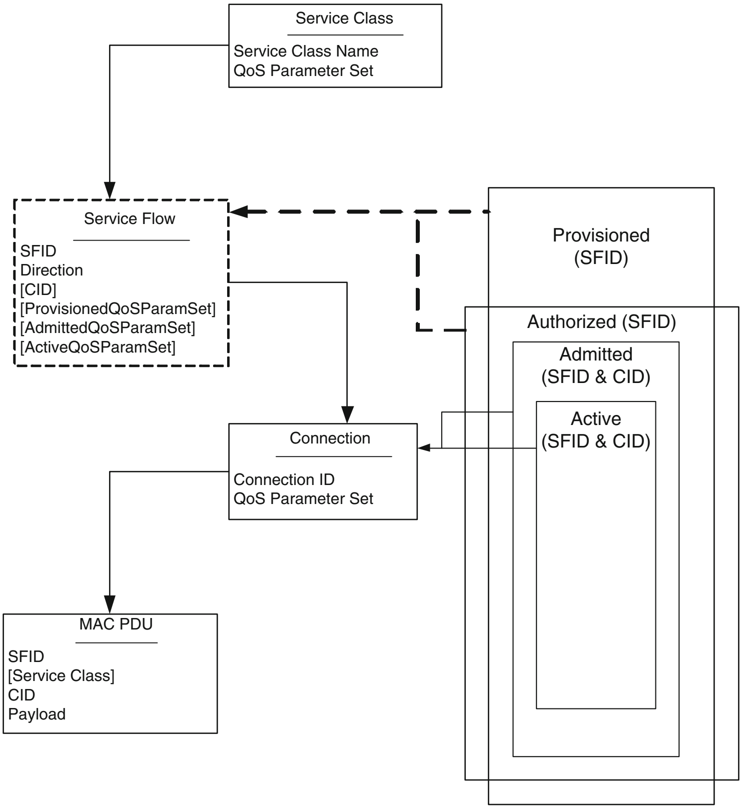


Fig. 9.10 Theory of operation and parameter sets

Service flows are defined as unidirectional description of connections. There is a unique Service Flow IP (SFID) per flow. Service flow parameters are categorized into service classes, which allow operators to provision the MS with the Service Class Name. The parameters of the service class might change from BS to BS but no further change is needed in MS as long as the MS is provisioned with the same Service Class Name.

There are three types of service flows as seen in Fig. 9.10: provisioned, admitted, and active service flows. Provisioned service flows are the ones that are not immediately activated but a SFID is assigned. MS or BS can choose to activate the service flow; MS sends the SFID and the QoS Parameter Set in DSC-REQ message to request for activation and BS responds with a CID if the flow is authorized; BS activates by sending DSC-REQ message with CID to MS.

There are two types of authorization: static and dynamic. In the static authorization, authorization module has the information of all provisioned service flows, and authorized set and admitted set are equal. In the dynamic authorization there is a policy server that provides the authorization module with advance notice of upcoming admission and activation requests. Requests from MS are checked by the authorization module to ensure that the activation requests from an MS are a subset of the set provided by the policy server. BS fetches the provisioned QoS set for an MS during the network, entry and authorization module in the BS uses this information to authorize the dynamic flows that are a subset of provisioned QoS parameter set.

To conserve the bandwidth, active flows are a subset of admitted flows. Admitted flows are the first phase of activation. For example, it is relevant in call-waiting where a call is admitted but not active and another active call can use the resource instead. Another example is when there is a call, the appropriate codec or wireless modulation can dynamically be adjusted for better service. Therefore, there is more demand for bandwidth and less jitter. In this case, the adjustment is accepted up to the admitted QoS parameters. This two-tier approach is helpful to conserve the network resources, performing policy checks and preventing several service thefts.

There are three states for a service flow; service flow are created with DSA messages. Service flows are changed with DSC messages, and finally service flows are deleted with DSD messages.

Service addition/change/deletion for an uplink or a downlink flow can be initiated either from network/BS or MS. There is a three-way handshake for data delivery: REQ, RSP, and ACK messages need to be exchanged for proper signaling.

9.6 Sleep Mode for Mobility-Supporting MS

MS is a mobile device; consequently, power consumption should be handled meticulously. MS time to time negotiates for absence to minimize the power usage. Sleep mode, an optional capability for MS but mandatory for BS, intends to maintain the prenegotiated absence.

BS assigns a power saving class to an MS for each group of its connections, which has common demand properties. Best effort connection may use one class but two UGS connections can demand two different classes.

There are three power saving classes defined in the standard, which differ with respect to their parameters and means of activation and deactivation.

A MS can have more than one power saving class assigned. Each power saving class introduces sleep and listen intervals. Consolidation of the intervals results in nonoverlapping availability and unavailability intervals. Unavailability interval is a period of time when all power saving class intervals are in sleep state as seen in Fig. 9.11.

During unavailability interval MS may power down or perform other activities such as scanning and association with neighboring BSs, etc. If there is a connection that does not have any active power saving class, the MS is considered available on a permanent basis.

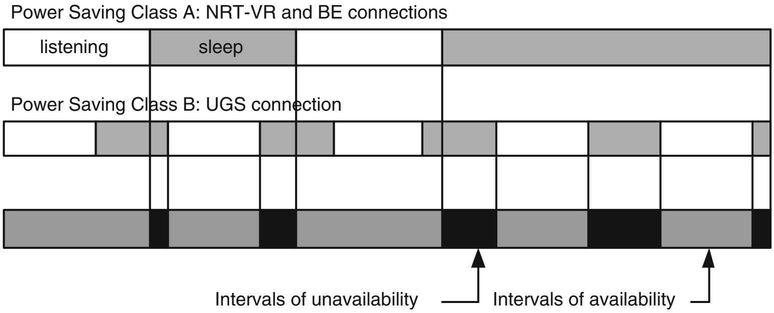

Power Saving Class A: NRT-VR and BE connections
listening
sleep
Power Saving Class B: UGS connection
Intervals of unavailability
Intervals of availability

Fig. 9.11 Example of unavailability interval in sleep mode

Power saving classes are activated by MOB_SLP_REQ/RSP message exchanges and power saving class is defined by the following parameters:

I: Initial-sleep window
FS: Final-sleep window base
L: Listening window
SE: Final-sleep window exponent
SF: Start frame number for first sleep window
T: Traffic-triggered wakening flag

Power saving classes are deactivated; if BS transmits a MAC SDU or fragment thereof over a connection belonging to the power saving class, MS transmits a bandwidth request with respect to connection belonging to the Power Saving Class, MS receives a MOB_TRF-IND message to indicate the presence of buffered traffic addressed to the MS.

9.6.1 Power Saving Class of Type I

Power saving class of type I is suitable for BE and NRT-VR traffic. In this type, sleep window (*SW*) is interleaved with fixed duration listening intervals and sleep window is doubled until it reaches the maximum value:

$$\mathrm{SW}(n+1) = \min(2 \times \mathrm{SW}(n), \mathrm{FS} \times 2^{\mathrm{SE}}). \tag{9.3}$$

Note that during listening intervals, MS receives all downlink transmission.

9.6.2 Power Saving Class of Type II

Type II is for UGS and RT-VR traffic where sleep window is fixed as initial window (I) and interleaved with listening windows (L) of fixed duration. Alternatively, type II can be defined and/or activated/deactivated by RNG-REQ/RSP message. During listening intervals, in addition to type I, MS can send or receive any data.

9.6.3 Power Saving Class of Type III

Power saving class is used for multicast connections as well as for management operations, for example, periodic ranging, DSx operations, neighbor advertisement message, etc. BS guesses time of the next portion of data in multicast service and allocates a sleep window. In the same way, BS may allocate a power saving class until the next periodic ranging interval. After the sleep window, MS becomes active for DL/UL transmission and BS may choose to reactivate the power saving class. RNG-REQ/RSP messages can be used for de/activation. Type III is executed once and inactive afterward. Sleep window is specified as base/exponent.

9.6.4 Periodic ranging in sleep mode

BS can interfere the listening window of the MS by allocating a UL transmission opportunity for periodic ranging or deactivate at least one power saving class to keep MS in active state, or let the MS know the next periodic ranging interval with *Next Periodic Ranging TLV*[3] in last successful RNG-RSP. If *Next Periodic Ranging TLV* is set to zero, MS resumes normal operation and all power saving classes are deactivated.

9.7 Handover

IEEE 802.16e-2005 defines two types of handover: *break-before-make* and *make-before-break*. The former one is a default scheme, which is mandatory in the standard. However, there are two optional schemes introduced for the latter. Both handover schemes require information about the neighboring BSs. This is obtained through scanning procedure.

9.7.1 Scanning

Information about neighboring BSs is advertised by the serving BS via MOB_NBR-ADV message. This message sent in DCD/UCD provides the channel information of neighboring base stations.

MS scans the neighboring BSs in order to detect the possible target BSs for handover. Time intervals for scanning are allocated to MS by BS. Also, MS may request an allocation of a group of scanning intervals with interleaving intervals of normal operation using the MOB_SCN-REQ/RSP in which MS indicates the estimated

[3] information is encoded as a Type Length Value element

duration of time it requires for the scan and also list of neighbor BSs to be considered for the scan. During the scanning period, serving BS buffers the incoming data and transmits either after the scanning interval or during any interleaving interval.

9.7.2 Association Procedure

Association procedure is introduced to assist scanning in order to accelerate the handover procedure. To decide for handover, MS or BS needs to learn from the prospective handover targets handover-related information such as the ranging parameters, service availability information, etc. There are three levels of association procedure introduced in the standard:

Level 0 – *Scan/association is without coordination:* MS ranges with the target BS using contention-based ranging allocations. Target BS after receiving the ranging code sends RNG-RSP with "success" and provides uplink allocation for the MS to transmit RNG-REQ message with related association ranging parameters.

Level 1 – *Association with coordination:* Serving BS coordinates the association procedure with the neighboring BSs. Neighboring BS allocates a ranging region at a predefined time (rendezvous time) in terms of frame number and assigns a unique CDMA code and transmission opportunity within the said region. The same code and transmission opportunity cannot be assigned to more than one MS. If MS can not obtain the UL-MAP at the first time immediately following the rendezvous time, MS performs *Level* 0 association.

Level 2 – *Network-assisted association reporting:* Serving BS prepares the neighboring BSs as in *Level* 1 but in this case MS only sends the CDMA code and does not wait for the RNG-RSP. Information of the RNG-RSP is sent to serving BS over the backbone. Serving BS sends this information to MS with MOB_ASC_REPORT message. If coordinated ranging is not successful at the first time immediately after rendezvous time, *Level* 0 is performed.

Handover decision algorithm utilizes all these information obtained through scanning and decides for handover, if needed. HO decision determines an ordered list of target BSs and triggers HO control module to initiate HO process.

9.7.3 HO Process

HO process depicted in Fig. 9.12 has two steps: preparation and action. Preparation is responsible for initiation of HO and preparation of the MS and network for handover. Handover initiation can be performed by either MS or BS or network. It

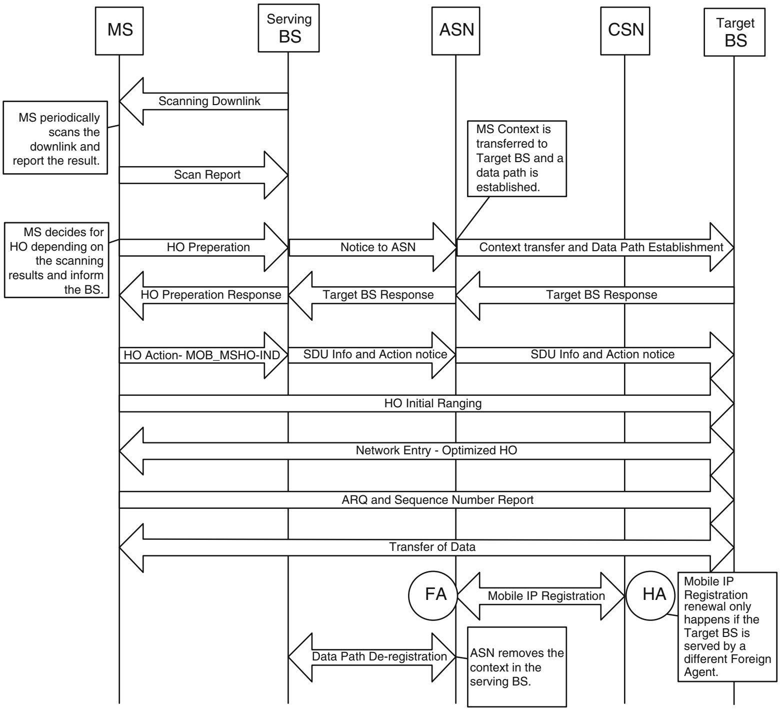

MS
Serving BS
ASN
CSN
Target BS
Scanning Downlink
MS periodically scans the downlink and report the result.
Scan Report
MS Context is transferred to Target BS and a data path is established.
MS decides for HO depending on the scanning results and inform the BS.
HO Preperation
Notice to ASN
Context transfer and Data Path Establishment
HO Preperation Response
Target BS Response
Target BS Response
HO Action- MOB_MSHO-IND
SDU Info and Action notice
SDU Info and Action notice
HO Initial Ranging
Network Entry - Optimized HO
ARQ and Sequence Number Report
Transfer of Data
FA
Mobile IP Registration
HA
Mobile IP Registration renewal only happens if the Target BS is served by a different Foreign Agent.
Data Path De-registration
ASN removes the context in the serving BS.

Fig. 9.12 Handover flow chart for mobile-initiated handover where handover controller resides in base station

depends in which entity handover decision algorithm resides. Action phase is responsible to perform necessary arrangement in order to seamlessly shift the MS from serving BS to target BS.

9.7.3.1 HO Preparation

If MS initiates the handover, it transmits MOB_MSHO-REQ message, which is acknowledged by BS with MOB_BSHO-RSP. If the handover initiation comes from the network side, BS transmits MOB_BSHO-REQ to inform MS for the handover. There are several rules associated with this preparation:

- Mobile initiation has higher priority than BS and network initiation.
- MS and BS can propose more than one target BS based on the MS's possible performance at the target BS.
- Serving BS can coordinate with target BS to allocate dedicated ranging transmission to skip CDMA ranging.
- Network-assisted HO allows MS to handover to any BS without any notification to the serving BS.

- MS can reject the HO if it does not prefer to handover to the selected BS by its serving BS. Then, serving BS alters the list.
- BS can force MS to conduct handover. MS is required to handover but still have option not to perform handover to the selected BS but others.
- MS sends MOB_HO-IND message to the serving BS to indicate the beginning of HO action. This is the last communication with its serving BS. However, serving BS keeps the context of MS until resource retain timer expires.
- MS can cancel the handover anytime and can indicate the serving BS with HO-IND message with HO cancel option set. Serving BS resumes normal operation.
- MS includes serving *BSID TLV* and *Ranging Purpose Indication TLV* in the RNG-REQ message sent to target BS. Target BS identifies the process if there is a handover with these information and assigns Basic and Primary CIDs.
- MS or BS can detect a drop, defined as the situation where an MS has stopped communicating with its serving BS. MS ranges with the target BS and target BS retrieves the context of the MS from the network. Serving BS reacts as if a MOB_HO-IND message has been received.

9.7.3.2 HO Action

HO action phase is similar to network entry procedure where a MS communicates first time with a BS (target). However, this process can be shortened by the target BS's possession of MS information from the backbone network. Hence, some steps can be skipped and right after ranging, target BS can send REG-RSP (or SBC-RSP) or RNG-RSP with *registration TLV* items.

If some steps are skipped this is indicated in the *HO Optimization TLV*, which is sent in the RNG-RSP message of target BS. When MS finishes the reentry procedure, MS sends a Bandwidth Request header with zero bandwidth request. Target BS can notify MS about pending MS downlink data via RNG-RSP message. The target BS can forward the data to the MS and then reestablish the IP connectivity.

To reduce the packet losses in the handover, data integrity mechanism is provided for ARQ-enabled and SDU_SN-enabled connections. For ARQ-enabled flows, the ARQ block sequence number is already available at the MS. For ARQ-disabled flows, there is a SDU_SN number, which is a virtual SDU number sent by the BS intermittently. When this number is not sent by BS, MS increments the number every time it receives data for that particular flow. MS sends the SN Report MAC header to inform the target BS about the SDU sequence numbers for each flow when it switches over. This way target BS compares the buffered packets with these numbers in order to filter the ones that have already been received by the MS.

9.7.4 Soft Handover

IEEE 802.16e standard supports two modes of soft handover (*make-before-break*) mechanism: Macro Diversity Handover (MDHO) and Fast BS Switching (FBSS).

MDHO allows MS to transmit and receive from multiple BSs within a *diversity set* at the same time. FBSS allows MS to receive/transmit data from/to the Anchor BS that is selected within the diversity set. One of the BSs in the list is selected as the Anchor BS and can be altered in each frame.

Diversity set is a list of BSs that are already identified as a target BS for that particular MS. This set is updated according to mobility pattern and load condition of MS. MS sends MOB_MSHO-REQ to drop the serving BS from the diversity set if long-term CINR of a serving BS is less than *H_Delete Threshold*. Also, it adds a neighbor BS if long-term CINR of neighbor BS is higher than *H_Add Threshold*. These threshold values are sent by BS.

The BSs involving in soft handover should have the same set of CIDs for the connections that are established with the MS. The BS may assign a new set of CIDs to the MS during diversity set update through MOB_BSHO-REQ message.

Note that soft handover requires stringent time synchronization among BSs. The frames sent by the BSs in diversity set should be within prefix interval. As a result, BSs should have synchronized frame structure and frequency assignment.

9.7.4.1 MDHO: Macrodiversity Handover

MDHO exploits transmit diversity in uplink and downlink by transmitting and receiving from multiple BSs at the same time. For downlink MDHO, multiple BSs transmit synchronously the downlink data such that the diversity combining can be performed by the MS. For uplink MDHO, the transmission from an MS is received by multiple BSs such that selection diversity can be performed.

Control information transmitted in DL-MAP, UL-MAP, and FCH is handled with two different methods. First method allows MS to receive the control information only from the Anchor BS, which has the allocation information for the non-Anchor BSs. Second method allows MS to receive the control information from all BSs in the diversity set.

9.7.4.2 FBSS: Fast Base Station Switching

In FBSS, MS communicates only with the Anchor BS for management and traffic connections. However, transition from one Anchor BS to another is performed smoothly without invocation of HO procedure.

FBSS operation illustrated in Fig. 9.13 requires the time axis slotted by an ASR (Anchor Switch Reporting) slot that is M frame long. A switching period is introduced with duration L ASR slots ($L \times M$ frames). L can be configured with DCD to accommodate certain processes before switching.

FBSS procedure is described as follows:

- First ASR Slot: The MS detects a BS in the diversity set with better signal strength.

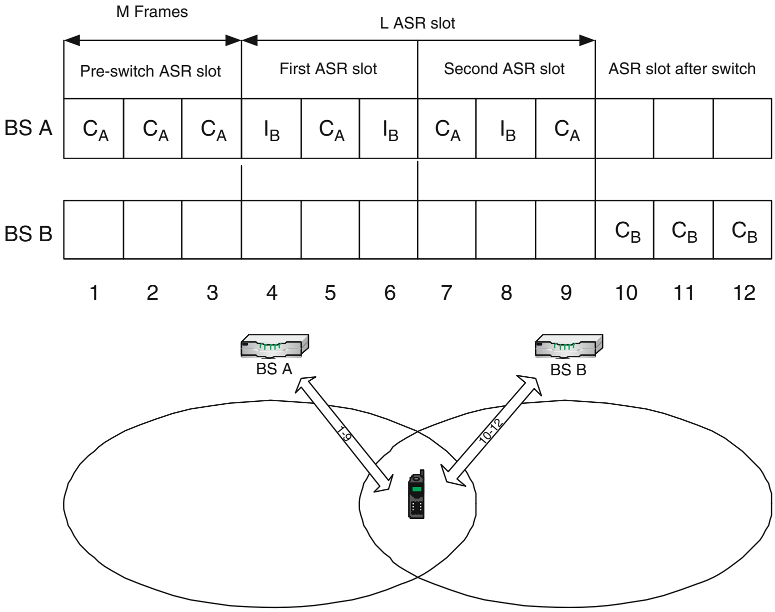


Fig. 9.13 An example of fast base station switching

- Second ASR Slot: The MS sends the Anchor BS switch indicator at the beginning and starts a switching timer with value equal to *L*.
- Second – L^{th} ASR Slot: The MS keeps sending the anchor switch indicator through CQICH allocated by the current Anchor BS. Anchor BS may send the *Anchor_Switch_IE* prior to the expiry of the switching timer to do one of the following: acknowledge the new Anchor BS, specify new action time, or cancel switching event. If the MS does not receive any indication, MS switches to the new Anchor BS after the L^{th} ASR Slot.
- Prior to expiry, the MS reports CQI and anchor switch indication on alternate frames.
- After the expiry, if the existing Anchor BS receives CQI transmission, the existing Anchor BS assumes that the MS has canceled the switch.
- After the switch, new Anchor BS allocates CQICH for MS to send CQI. If MS does not receive a CQICH allocation for a duration equal to the switching period, the MS requests CQICH by transmitting CQICH allocation request header.

FBSS is simpler as compared to MDHO since there is no need for diversity combining. On the other hand, in the network side it requires more sophisticated data integrity as compared to MDHO.

9.8 MBS: Multicast Broadcast Service

Multicast and broadcast services are designed for multimedia-related service flows for downlink operation. A service flow that is multicast or broadcast by BS can be shared by multiple MSs. The standard introduces two prong ways: Single-BS MBS and Multi-BS MBS. If a service flow is shared only within a cell it is a single-BS operation. If neighboring BSs also broadcast the same service flow at the same location in the same frame number, then it is a Multi-BS operation. Remember that a service flow is transported via a CID. For single-BS, that CID is any available traffic CID and shared by all MSs within the cell. For multi-BS, the multicast CID is used, which shall be the same for all BSs on the same channel that participates in the connection. Multi-BS access improves the reliability by benefiting from diversity. However, it requires time synchronization among BSs as well as scheduling the data in the same coordinates of the frame.

MBS_Zone is defined for Multi-BS operation to allow MS receive service without registering to the BS from which it receives the transmission. MBS_Zone is advertised in DCD message to identify the BSs, which use the same CID and corresponding SA. As a result, an MS coming from idle mode into the same MBS_ZONE does not need to reestablish the flow.

Also, during creation of multicast connection via DSA-REQ/RSP messages, an MBS connection for multiple MBS contents can be established by using an *MBS Contents Identifier TLV* encoding. As a result, an MS receives multiple MBS messages for an MBS connection and identifies them by *Logical Channel ID*, which belongs to a Multicast CID.

Note also that MBS connections are ARQ-disabled and MSB encryption is performed using Globally Transport Encryption Keys (GTEKs). However, service flows are encrypted at the application layer or MAC layer or both by AES-CTR[4]. This prevents unauthorized access to multicast and broadcast content. An MBS_MAP_IE is inserted to DL-MAP to introduce power efficient operation for an MS when receiving the MBS data since this IE directs the MS to the MBS allocation.

9.9 Idle Mode and Paging

When there is no activity, MS goes to idle mode. Idle mode is designed to remove the active requirement for HO, and all normal operation requirements. MS only performs scanning at discrete intervals while it traverses a coverage area constructed by multiple BSs.

Paging is designed to direct traffic toward MS when MS is in idle mode. Paging is responsible to locate and wake up the MS. A number of adjacent BSs constitute a paging group as can be seen in Fig. 9.14. A BS can be part of more than one paging

[4] Defined in NIST Special Publication 800-38A, FIPS 197.

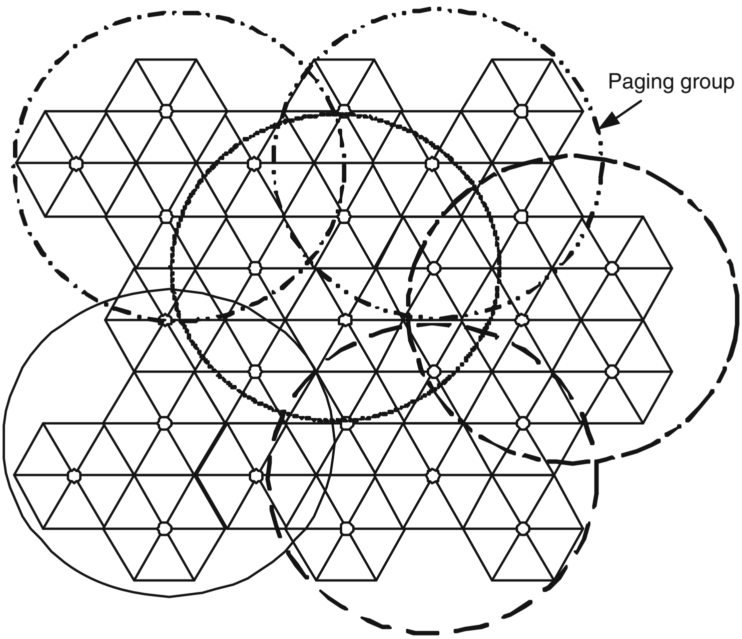


Fig. 9.14 Paging group

group in order to provide continuous coverage. When MS crosses a new paging group, it has to report its new paging group to the network. If paging group size is small, MS reports more often and consumes more wireless bandwidth. If paging group size is large, then it is hard to locate the MS. As a result, size of the paging group determination lies on this tradeoff.

Idle mode and paging procedure are described here:

- *MS idle mode initiation:* Idle mode initiation begins either with DREG-REQ message, sent by MS or DREG-CMD message, sent by BS. Paging Controller residing in the serving BS retains certain MS context to assist MS in the future reentry from idle mode. MS and *Paging Controller* has a timer to prompt MS idle mode *location update* activity and demonstrates MS continued network presence to revalidate Paging Controller about the retention of MS service and operational information. When MS enters idle mode, ARQ state information and parameters between MS and BS are removed.
- *Unavailable/listening interval:* MS selects a preferred BS, which is a neighbor BS with the best DL properties. The MS decodes the DCD and DL-MAP for the preferred BS to determine the next paging interval for the preferred BS. The duration until the next regular BS paging interval is the *unavailable interval* for the MS. The MS powers down, scans neighbor BSs, reselects a preferred BS, and conducts ranging but does not guarantee availability to any BS for DL traffic.

The MS scans, decodes the DCD and DL-MAP, and begins decoding any paging message during the entire BS paging interval.

- *BS broadcast paging message:* MS is notified by a broadcast message to indicate either the presence of DL traffic pending or requirement for network entry or to poll the MS and requests a location update without requiring a full network entry. A paging message is transmitted during the *listening interval* for any MS that needs paging and a ranging is expected. If BS does not receive RNG-REQ signal from MS, after it transmits the paging signal up to "Paging Retry Count," BS determines and informs the network that the MS is unavailable. MS can terminate idle mode and reenter the network if it decodes a BS Broadcast Paging with its own MS MAC Address hash. MS performs location update at will or if any one of the four conditions is met:

  – Paging group change, which is detected by PG_ID in DCD or MOB_PAG-ADV during the listening interval
  – Periodically prior to the expiration of the Idle Mode Timer
  – As part of its power-down procedure
  – MAC Hash Skip Threshold

  Location update can be secure or unsecure. In secure update, MS includes HMAC/CMAC Tuple in the RNG-REQ. If MS and BS do not share current and valid security context, they follow location update with network re-entry. MS initiates network re-entry with the target BS by sending RNG-REQ with Paging Controller ID TLVs. If target BS does not have the MS context, it requests it from the network.

9.10 Summary

We described the MAC layer of WiMAX. The MAC is an interface in between PHY and network architecture. The PHY features and functions such as construction of PDUs, ARQ, resource allocation, etc described in the previous chapter are complemented in this chapter. Also, the protocols and features described in this chapter are complemented in the next chapter at network level such as security, QoS, mobility management, and idle mode and paging.

References

1. *IEEE Standard 802.16-2004, Part 16: Air interface for fixed broadband wireless access systems,* June 2004.
2. *IEEE Standard 802.16e-2005, Part 16: Air interface for fixed and mobile broadband wireless access systems,* December 2005.

Chapter 10
WiMAX Network Layer

10.1 Introduction

This chapter describes the networking protocols that complement the end-to-end functions of WiMAX technology together with PHY and MAC signaling of IEEE 802.16 standard, described in the previous two chapters, respectively. WiMAX Forum, which is constituted by the companies, formed Working Groups[1] to define the networking protocols and facilitate the adoption of WiMAX technology. Networking Working Group is responsible to develop the networking architecture in conjunction with other groups, especially with Service Provider Working Group that basically defines the desired level of standards by the operators.

This chapter is adapted from WiMAX Forum documents with permission. The topics presented in this chapter are derived from published Stage 2 and Stage 3 specifications for Release 1 Version 1.2. Also, this chapter contains information about ongoing developments of Release 1.5 in brief. We advice the user that although information from WiMAX Forum Stage 2 and Stage 3 documents is up to date, final versions of these specifications may contain minor changes.

In this chapter, you will find the first introduction to IP-based access network. In the later chapters of this book, which cover other IP-based OFDMA technologies, one will notice the similarities in terms of functionalities as well as entities. As a

[1]
- Application Working Group or Application Architecture Task Group
- Application Business Task Group
- Certification Working Group
- Evolutionary Technical Working Group
- Global Roaming Working Group
- Marketing Working Group
- Network Working Group
- Network Interoperability Task Group
- Regulatory Working Group
- Service Provider Working Group
- Technical Working Group

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,

DOI: 10.1007/978-0-387-68192-4_10, © Springer Science+Business Media LLC 2009

result, IP-based access network is explained in more detail here than in the following chapters to avoid duplication as much as possible.

First, we introduce the design constraints. Although design constraints explained in the next section are specifically for WiMAX architecture, these are also applicable to other IP-based networks as well. Then, we continue the discussion with reference model and functional entities adopted by WiMAX network. Finally, we end this chapter with a brief summary of upcoming features after discussing the network protocols and functions such as network entry, IP addressing, AAA, QoS, and mobility.

10.2 Design Constraints

The WiMAX NWG defines the end-to-end architecture in a three-stage standard: Stage 1 is for use case scenarios and service requirements and is defined along with Service Provider Working Group specifications; Stage 2 describes the architecture; Stage 3 details the architecture. WiMAX design principles include the following:

- The architecture shall be decomposed into functions and well-defined reference points between functional entities for multivendor interoperability.
- The architecture shall provide modularity and flexibility in deployment. Multiple types of decomposition topologies may coexist such as distributed, centralized, and hybrid.
- The architecture shall support fixed, nomadic, portable, and mobile operation and evolution paths to full mobility.
- The architecture shall support decomposition of access networks and connectivity networks. The access network is radio-agnostic, and connectivity networks provide IP connectivity.
- The architecture shall support sharing of the network with a variety of business models:
  - Network Access Provider (NAP) owns the network and operations.
  - Network Service Provider (NSP) owns the subscriber and provides service. NSPs share the NAP or a NSP uses multiple NAPs.
  - Application Service Provider (ASP) provides application services.
- The architecture shall support internetworking with 3GPP, 3GPP2, WiFi, and wireline networks defined with IETF protocols.

10.3 Network Reference Model

Network Reference Model, seen in Fig. 10.1, defines functional entities and reference points regarding interoperability between vendors. Functional entities are grouped into three sets: Mobile WiMAX Subscriber (MS), Access Service Network

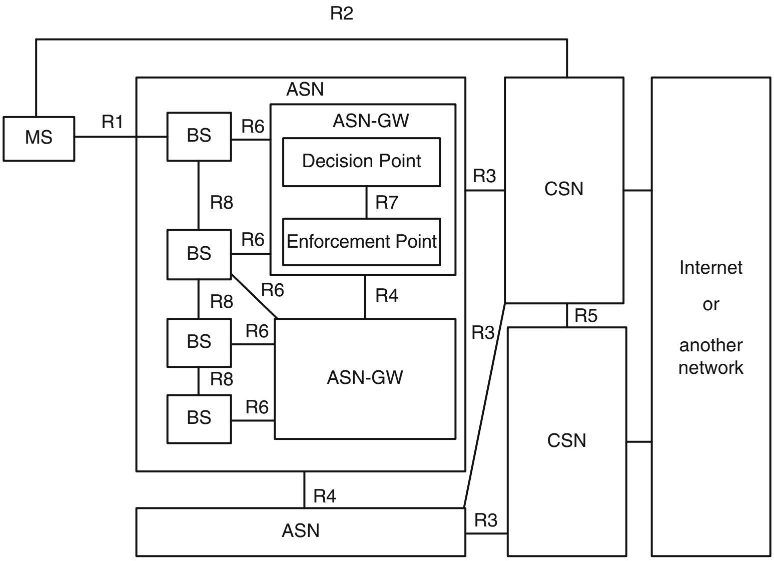


Fig. 10.1 Network reference model (© WiMAX Forum 2005–2007)

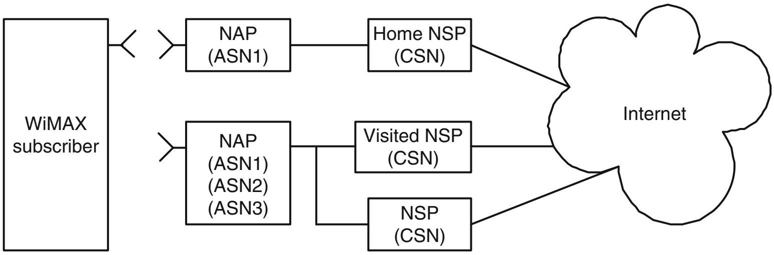


Fig. 10.2 Business relationship between WiMAX subscriber, NAP, and NSPs

(ASN), and Connectivity Service Network (CSN), where OFDMA and IP define the demarcation between the sets; ASN is radio-agnostic and responsible for assisting MS to maintain the uninterrupted connectivity of OFDMA air link during mobility and idle mode; CSN provides set of network functions to provide IP connectivity services to the subscriber.

Figure 10.2 shows the business relationship between functional entities. ASN is owned by NAP and CSN is owned by NSP. A NAP can be shared by more than one NSP and a NAP can have more than one ASN. A subscriber has a home NSP where relation in between is established by a service level agreement (SLA). Between NSP and NAP, there is a contractual agreement. And, roaming of subscriber is performed by establishing roaming agreements between visited and home NSPs.

10.4 ASN: Access Service Network

ASN is responsible to provide Layer-2 connectivity to MS and transfers the control/data messages to the MS coming from CSN. ASN comprises one or more BSs and one or more ASN Gateways (ASN-GWs). MS's connectivity and continuity to WiMAX network with optimized performance requires following functionalities from ASN:

- Network discovery and selection of the WiMAX subscriber's preferred NSP
- Network entry with IEEE 802.16e-2005-based Layer-2 connectivity with WiMAX MS
- Relay function for establishing Layer-3 connectivity (IP address allocation) with a WiMAX MS,
- Radio Resource Management
- Multicast and Broadcast Control
- ASN anchored mobility
- Foreign Agent functionality for CSN anchored mobility
- ASN–CSN tunneling
- Paging and Location Management
- Accounting assistance
- Data forwarding
- Service flow authorization
- Quality of Service
- Admission Control and Policing

These functions can be distributed among BS and ASN-GW according to profile definitions defined in Release 1.0 of WiMAX Forum NWG Standard. There are three defined profiles that map a functionality to an entity: Profile A, Profile B, and Profile C where Fig. 10.3 illustrates Profile C functional distribution. Profile A[2] gives handover and radio resource management control to ASN-GW and leaves handover and radio resource management agent to BS. Both profiles require exposed intra-ASN interface between BS and ASN-GW. On the other hand, Profile B is characterized as unexposed intra-ASN interface where any combination of mapping is possible. All profiles require interoperability with CSNs and other ASNs via respective *reference points*. Let us first look at the entities residing in an ASN.

10.4.1 BS: Base Station

The WiMAX BS is a logical entity that implements an interface to air link and IP network. The BS embodies IEEE 802.16e-2005 PHY and MAC layers as well

[2] WiMAX Forum removed Profile A from the upcoming releases. Also, WiMAX Forum NWG has recently selected Profile C as the only ASN functional decomposition profile but allowed integrated ASN models.

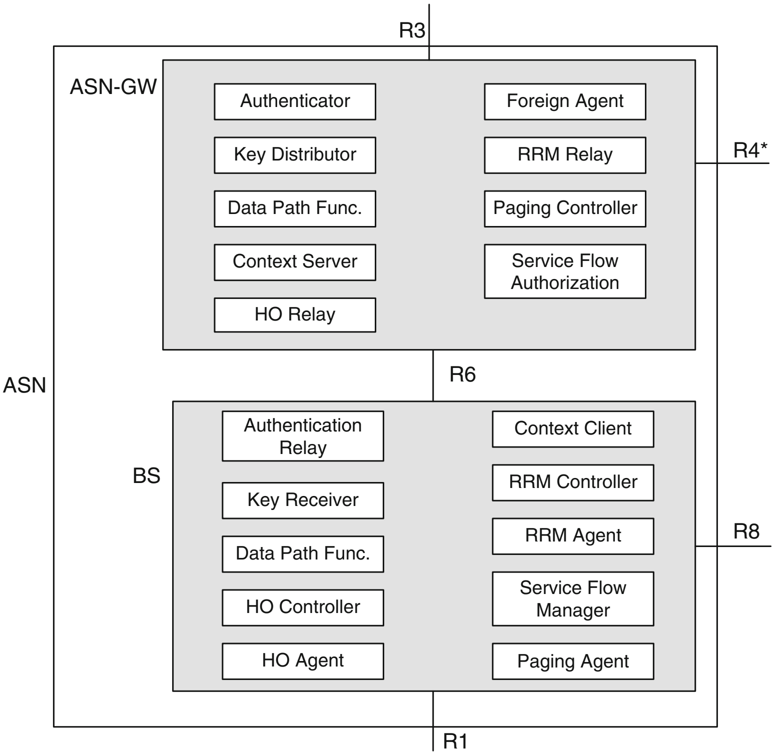


Fig. 10.3 Profile C functional architecture (© WiMAX Forum 2005–2007)

as one or more ASN functions to facilitate communication to ASN-GW and other BSs. The IEEE 802.16e-2005 BS instance represents one sector with one frequency assignment and a single BS may have a connectivity to more than one ASN-GW for load balancing or redundancy or both. A physical BS may have multiple BSs since BS is defined as a logical entity. The key component of BS is scheduler, which is responsible to allocate uplink and downlink resources in the air link.

10.4.2 ASN-GW: Access Service Network - Gateway

The ASN Gateway (ASN-GW) is a logical entity that aggregates the control plane and security functions as a Decision Point (DP). The ASN-GW may also perform bearer plane routing or bridging function as Enforcement Point (EP). Interface between DP and EP is unexposed and DP can be shared among BSs.

In Profile C, ASN-GW DP hosts functions listed in Fig. 10.3. ASN-GW DP holds radio-agnostic control. ASN-GW may have the authenticator and key distributor to implement AAA framework along with AAA relay. AAA framework verifies the user credentials during network re/entry with EAP authentication and creates security context with keys that are shared in BS and MS. AAA framework is also responsible for accounting. ASN-GW is also responsible for profile management together with Policy Function of CSN. Profile management is responsible to retrieve flow information and QoS parameters in order to be received in service flow authorization and admission control. ASN-GW also creates a context per MS upon network entry. Context management stores information about MS such as profile information, security context, etc. This context is updated and exchanged during handover with target BS.

During handover, ASN-GW is responsible to switch the data path to target BS. ASN-GW also participates in data integrity to minimize the latency and packet loss. ASN-GW hosts Foreign Agent to assist layer-3 handover in order to communicate to the Home Agent if mobility is across ASN. Paging Controller and Location Register in ASN-GW assist paging and idle mode operation.

ASN-GW EP is responsible to map radio bearer to the IP network and perform packet filtering, tunneling, admission control and policing, QoS, and data forwarding. EP routing capability may include IPv4/v6 unicast/multicast routing protocols such as BGP, OSPFv2, PIM, IGMP, etc.

ASN-GW is a central point in the ASN and the outer interface to other networks as well as IP services. ASN-GW may utilize the provided information and decision set to offer network optimization.

10.5 CSN: Connectivity Service Network

Connectivity Service Network (CSN) is defined as a set of network functions that provide IP connectivity services to the WiMAX subscriber. A CSN may facilitate following functions:

- IP address allocation
- Internet access
- AAA proxy or server
- Policy and Admission Control based on user subscription profiles
- ASN–CSN tunneling support
- Billing and interoperator settlement
- Inter-CSN tunneling for roaming
- Inter-ASN mobility
- WiMAX services: Location-Based Services, peer-to-peer services, IMS services, lawful intercept, multicast broadcast services, etc.

These functions are distributed to AAA proxy/servers, Policy Function, Home Agent, Internetworking gateway, etc in CSN.

10.6 Reference Points

Reference points are defined to boost interoperability between vendors. Reference points are logical interfaces that define the communication signaling between two peer functional entities. Functional entities may be colocated or physically separated. Reference points shown in Fig. 10.1 are defined here:

R1: Reference point is between MS and BS and defined by the IEEE 802.16e-2005. R1 is authenticated, integrity and replay protected at the IEEE 802.16 MAC layer upon successful *Device Authentication*.

R2: Reference point is between MS and ASN-GW or CSN. This is a logical interface that uses the entities in between as pass through and is typically used for authentication, authorization, IP configuration, and mobility management. R2 may not have end-to-end secure channel and relay signaling and data traffic security assuming insecure lower layer.

R3: Reference point is between ASN and CSN. This reference point defines communication to AAA, Policy Function, and HA. R3 may not have an end-to-end secure channel, and the signaling and data traffic security should assume its own security assuming insecure lower layer. For example, MIPv4 may use authentication extensions; RADIUS may use authentication attributes, etc.

R4: Reference point is between ASNs. This reference point is used during the mobility across ASNs. R4 may have an end-to-end secure channel, including privacy with IPSec or SSL VPNs, etc.

R5: Reference point is between CSNs. This reference point is used when MS is in a visited network and needs to communicate to home network. R3 may not have an end-to-end secure channel, and the signaling and data traffic should assume its own security assuming insecure lower layer. For example, RADIUS may use authentication attributes.

R6: Reference point is between ASN and CSN. This reference point is the most used one. It is an exposed interface in Profiles A and C but unexposed in Profile B. It is used for control signaling and implements intra-ASN tunnels. R6 may have an end-to-end secure channel, including privacy with IPSec or SSL VPNs, etc.

R7: Reference point is unexposed and between data and control plane of ASN-GW.

R8: Reference point is between BSs. This interface is being designed to be used for signaling during handover and load balancing. R8 may have an end-to-end secure channel, including privacy with IPSec or SSL VPNs, etc.

10.7 Protocol Convergence Layer

The ASN is responsible to aggregate and forward control and data packets coming from MS and CSN, respectively. The MS data packets are transferred from BS to ASN-GW over *data paths*. ASN can support both IP and Ethernet packets through

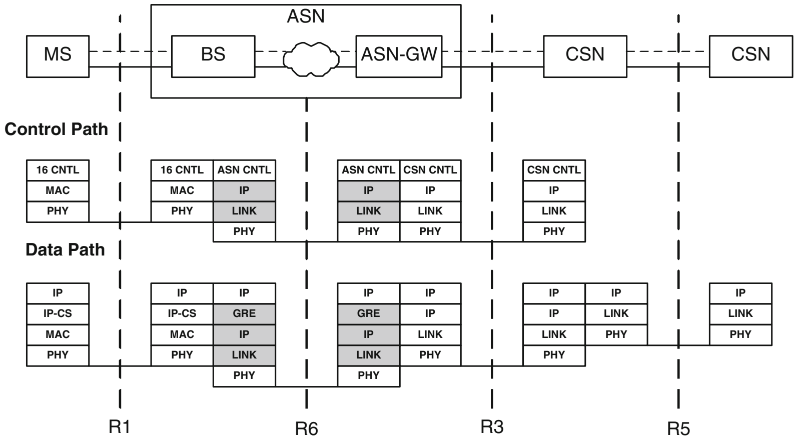


Fig. 10.4 IP-CS with routed ASN: If it is bridged ASN, then the shaded region would be replaced with an Ethernet layer (© WiMAX Forum 2005–2007)

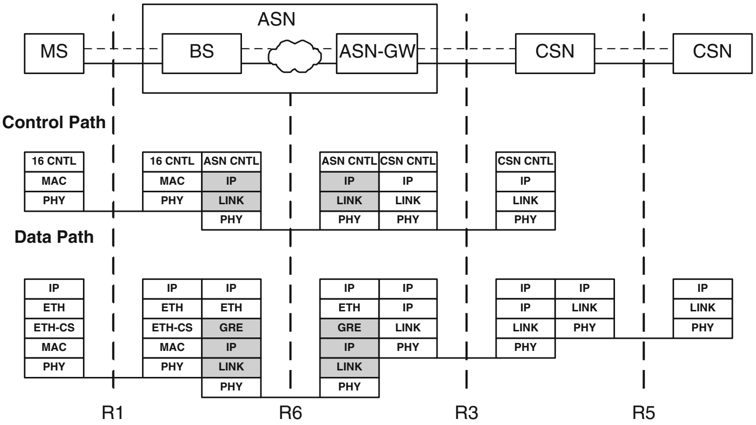

ASN
MS
BS
ASN-GW
CSN
CSN
Control Path
16 CNTL | MAC | PHY
16 CNTL | MAC | PHY
ASN CNTL | IP | LINK | PHY
ASN CNTL | IP | LINK | PHY
CSN CNTL | IP | LINK | PHY
CSN CNTL | IP | LINK | PHY
Data Path
IP | ETH | ETH-CS | MAC | PHY
IP | ETH | ETH-CS | MAC | PHY
IP | ETH | GRE | IP | LINK | PHY
IP | ETH | GRE | IP | LINK | PHY
IP | IP | LINK | PHY
IP | IP | LINK | PHY
IP | LINK | PHY
IP | LINK | PHY
R1
R6
R3
R5

Fig. 10.5 Ethernet-CS with routed ASN: If it is bridged ASN, then shaded region would not be needed. Notice that Ethernet packets can be relayed up to CSN with another GRE tunnel between ASN-GW and CSN to enable VLAN services (© WiMAX Forum 2005–2007)

routed or *bridged* ASN. IP packets may be transported through IP-CS or Ethernet-CS over IEEE 802.16e-2005.

IP-CS carries directly IP datagrams of 802.16 PDUs as seen in Fig. 10.4. Single or multiple hosts can reside behind MS where MS encapsulates IP datagrams from the IP host layer into 802.16 MAC frames for upstream over the R1 reference point. The BS encapsulates IP datagrams received from the ASN-GW IP router via R6 into 802.16-MAC frames for downstream over R1. Ethernet CS carries the IEEE 802.3 frames in the payload of 802.16 PDUs as seen in Fig. 10.5. Notice that these two convergence layers coexist with the same ASN but inter-CS handover is not supported.

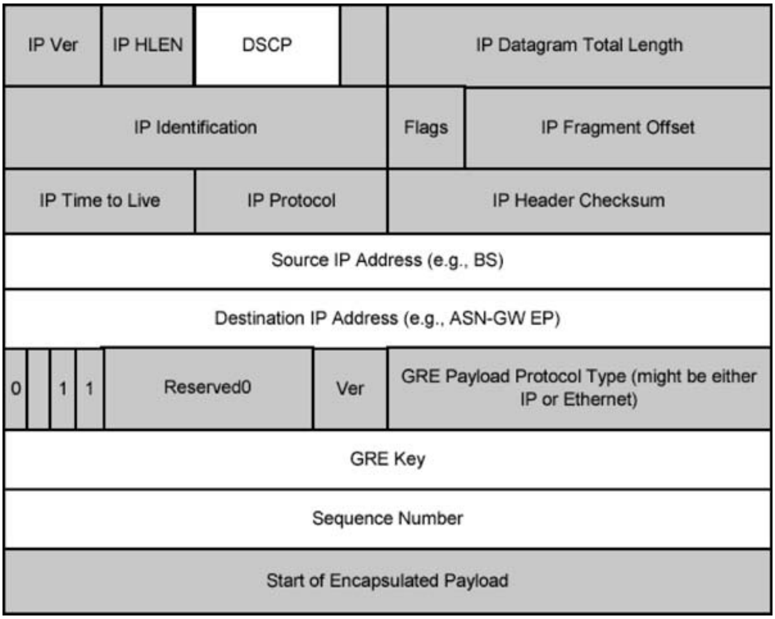


Fig. 10.6 GRE encapsulation

Also, note that Figs. 10.4 and 10.5 illustrate the routed-ASN where IP-in-IP encapsulation protocols are utilized in order to hide the IP address of the destination when routing takes place between BS and ASN-GW. In the downlink, ASN-GW creates a tunnel to BS that has a tunnel header representing the BS address as the destination address and ASN-GW address as the source address. BS strips off the tunnel header and forwards the packet to IEEE 802.16e-2005 interface. Another tunnel is created for uplink as well in which now BS creates the tunnel header toward ASN-GW.

Generic Routing Encapsulation (GRE) is used in WiMAX to create tunnels within ASN. GRE tunnels are created implicitly without control signaling. There can be one GRE tunnel per service flow or per MS or per BS, and each tunnel is identified by a tunnel ID. If there is a specific tunnel per service flow, then in ASN-GW during classification of the downstream traffic, the tunnel ID is one-to-one mapped to a SFID. But, if there is a tunnel per MS, then BS needs to classify the downstream traffic in order to map the flow to a particular CID (Fig. 10.6).

10.8 Network Discovery and Selection

A WiMAX subscriber in order to attach a WiMAX network needs to perform network discovery and selection. Any WiMAX subscriber, which could be nomadic, portable, or fully mobile, must perform following steps in order to get serviced from a WiMAX network:

- *NAP Discovery:* First, the subscriber needs to attach to a NAP. NAP is the owner of ASN, which basically provides the infrastructure for connection. Operator ID[3]-24 bit is embedded into the Base Station ID of DL-MAP. The subscriber detects the NAPs by scanning and decoding the DL-MAPs and selects the one that broadcasts the NSP list, which the subscriber has right to access.
- *NSP Discovery:* Then, the subscriber needs NSP to get IP connectivity. A unique NSP ID[4]-24 bit is also broadcasted. If more than one NSP uses the same NAP, then NSP list is provided to the subscriber from which the subscriber makes either an automatic or manual selection.
- *NSP Enumeration and Selection:* The subscriber either performs an automatic selection, which requires a dynamic information obtained within the area, or manual selection, which requires the user to exercise a provisioning procedure initially.
- *ASN Attachment*: After selection of NSP, the subscriber provides its identity and home NSP domain in the form of NAI (Network Access Identifier). From the realm portion of NAI, the ASN figures out the next AAA hop. The subscriber may provide a routing choice with decorated NAI as well if the home NSP is reachable through another NSP. For instance, "NSP4!user-name@NSP1.com" is a decorated NAI where NSP4 is reachable through NSP1.

10.9 IP Addressing

IPv4 addressing delivers the Point of Attachment (PoA) address to MS. DHCP is selected as the primary mechanism to assign dynamic IP address if MS does not support Client MIP. ASN implements a DCHP relay to pass through the DHCP signaling between MS and DHCP server. Alternate configurations are assigning IP address by a AAA server or Home Agent and delivering via DHCP. In this case, ASN implements a DHCP proxy to implement DHCP signaling between MS and ASN. MS IP address assignment is done via R1, R3, and R5, if applicable, interfaces. IP address may be allocated by the Address Allocation Server (AAS) either residing in the visited NSP or home NSP.

IPv6 addressing differs a little bit. MS maintains two IP addresses: Care-of-Address (CoA) and Home Address. The PoA can either be CoA whose scope is at the IPv6 AR (Access Router) or Home Address whose scope is at the MS's home agent. MIP6 MS can use either of the addresses for its IP sessions.

IPv6 AR resides in the ASN-GW and establishes a point-to-point link with MS as in Fig. 10.7. Either DHCPv6 [RFC3315] or stateless address autoconfiguration (SLAAC) is used to allocate CoA. Unlike DHCPv6 that employs stateful address configuration, SLAAC is stateless and performed between MS and AR after initial service flow (ISF) establishment. The MS sends router solicitation and also AR

[3] Operator ID is assigned as an IEEE 802.16 Operator ID by the IEEE Registration Authority.

[4] NSP ID allocation and administration are managed by the IEEE RAC.

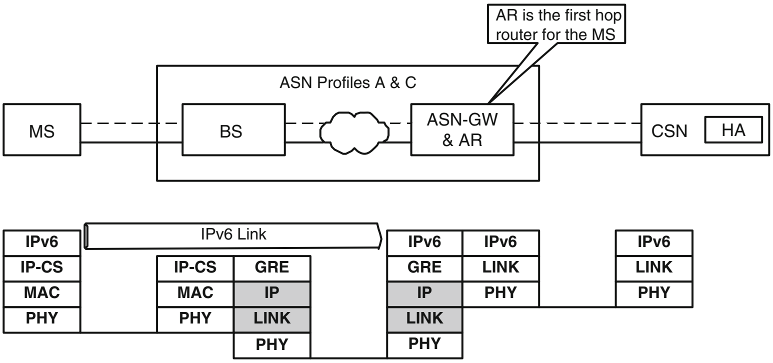


Fig. 10.7 IPv6 link model for Profiles A and C. Notice that in Profile B, the link between BS and the AR is unspecified

sends router advertisements, which include prefix(es) that enable MS to autoconfigure an address. MS performs Duplicate Address Detection (DAD) on the address it configures.

HoA for an MS is assigned with Stateless DHCP; home AAA sends to visited AAA the MIP6 bootstrap parameters that include the home agent address, the home link prefix, or the HoA in the stateless DHCPv6 server. After ISF, the MS sends DHCPv6 query to the DHCP server and receives the MIP6 bootstrap parameters. MS uses the HoA in the binding update to the HA. If HoA or home link prefix are not given by home AAA, MS can use an unspecified address and HA assigns HoA to the MS in the binding ACK.

10.10 AAA Framework

The AAA specifies the protocols and procedures for authentication, authorization, and accounting associated with the user and subscribed services across different access technologies. The AAA framework provides the following:

- *Authentication services:* AAA framework is responsible for authentication services such as device and/or user authentication.
- *Authorization services:* AAA framework is responsible for authorization services, which include delivery of information to configure the session for access, mobility, QoS, and other applications.
- *Accounting services:* AAA framework is responsible for accounting services, which include prepaid, postpaid, and hotlining activity.

The WiMAX AAA framework considers several functional requirements set by the WiMAX Forum. The AAA shall do the following:

- Support global roaming across WiMAX operators and roaming between home and visited NSPs

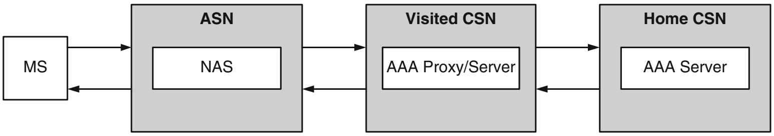


Fig. 10.8 Greenfield Roaming AAA Framework (© WiMAX Forum 2005–2007)

- Support EAP-based authentication mechanisms that include but are not limited to passwords, shared secrets, subscriber identity module (SIM), universal subscriber identity module (USIM), universal integrated circuit card (UICC), removable user module (RUIM), and X.509 digital certificates
- Be based over RADIUS or DIAMETER protocols
- Consider MS as "Supplicant," ASN as an "Authenticator", and AAA server as an "Authentication Server" and support PKMv2 for MS authorization, user authentication, and mutual authentication between MS and the NSP.
- Accommodate both Mobile IPv4 and Mobile IPv6.

The AAA framework [RFC2904] can be deployed in three ways: agent sequence/model, pull sequence/model, and push sequence/model. The pull model is recommended within WiMAX networks. A roaming AAA framework is depicted in Fig. 10.8 where MS sends request to Network Access Server, which forwards the request to the home AAA server via AAA proxy in the visited CSN. Routing of AAA packets over RADIUS is determined by NAI that is used for user and device authentication. Outer identity of NAI is used to route the packet to EAP authentication server and inner identity is used to identify the user, or authentication credentials. AAA employs hop-by-hop routing and RADIUS provides integrity protection, privacy, and protection against replay attacks. RADIUS may be protected with IPSec. NAS architecture in the ASN consists of the following functional entities: Authenticator, Authentication Relay, Key Distributor, and Key Receiver together with the prepaid client, hot-line device, AAA client, and accounting client. Two modes of operation are allowed: integrated or standalone. The integrated mode colocates those four entities in the BS; on the other hand, the standalone mode only resides Key Receiver and Authentication Relay in the BS, and the rest in the ASN-GW. This is applicable for Profile A and C of WiMAX NWG Release 1.

10.10.1 Authentication and Authorization Protocols

PKMv2 with Extensible Authentication Protocol (EAP) is specified by the IEEE 802.16-2004, and IEEE 802.16e-2005[5] to provide device and user authentication. Over-the-air user authentication is provided with PKMv2 protocol, and EAP messages are sent to Authenticator via Authenticator Relay. The Authenticator encapsulates the EAP messages and forwards them to AAA server via AAA proxies in

[5] Another standard is PKMv1, which only supports device authentication.

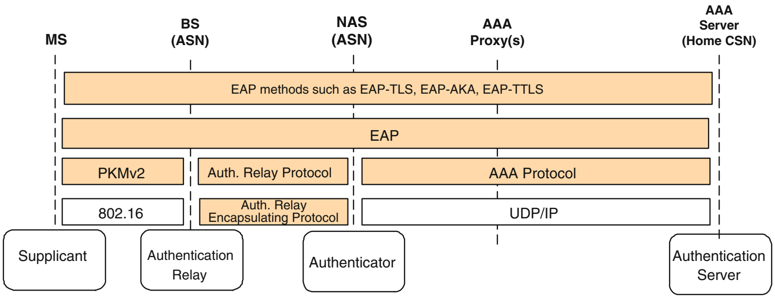


Fig. 10.9 PKMv2 User Authentication Protocols (© WiMAX Forum 2005–2007)

the CSN as seen in Fig. 10.9. For device authentication, EAP methods must generate Master Session Key (MSK) and Extended Master Session Key (EMSK) keys. Device authentication requirement is checked by MS from home CSN during provisioning process. If both user and device authentication need to be performed, there are two prong ways: EAP tunneling or combined authentication. Tunneling uses outer EAP for device, and inner EAP for user authentication over single EAP. Combined authentication still uses single EAP but uses a combined identity where a successful authentication between MS and home CSN insures both device and user authentication (Fig. 10.10).

Authentication and authorization are performed in network entry whose flow chart is depicted in Fig. 10.11. During capabilities exchange via SBC REQ/RSP messages, PKM version, PKMv2 security capabilities, and authorization policies are negotiated including the requirement for device authentication. Authenticator is notified and EAP exchange starts with EAP-Identity request. EAP exchange creates MSK and EMSK in the MS and the home AAA server. This MSK is transferred to the Authenticator (Key Distributor) from AAA via RADIUS but EMSK is retained in the AAA to be used in the derivation of mobility keys. From the MSK, both the MS and the authenticator generate PMK (Pairwise Master Key) as seen in Fig. 10.12. Authentication Key (AK) is derived from PMK in MS and the Authenticator. Key Distributor creates AK ID for <MS,BS> pair and delivers the AK and its context to the Key Receiver residing in the BS. From AK, various keys are derived, specified by IEEE 802.16e as seen in Fig. 10.12. During context transfer due to handover, keying materials including AK, CMAC_KEY_COUNT, AK Sequence Number, AKID, AK lifetime, EIK (EAP Integrity Key), etc., are transferred to new BS.

Security Association (SA) based on AK is verified through PKMv2 three-way handshake between MS and the BS. After the three-way handshake, MS and BS start using the AK for the protection of MAC messages. Three-way handshake starts with the BS transmitting *PKMv2 SA TEK Challenge*, which contains the AK and BS Random, a random number. The MS validates the AK and responds with *PKMv2 SA TEK REQ* with a random number. This also includes a request for SA descriptors identifying primary and static SAs, and group SAs (GSA). The BS responds with

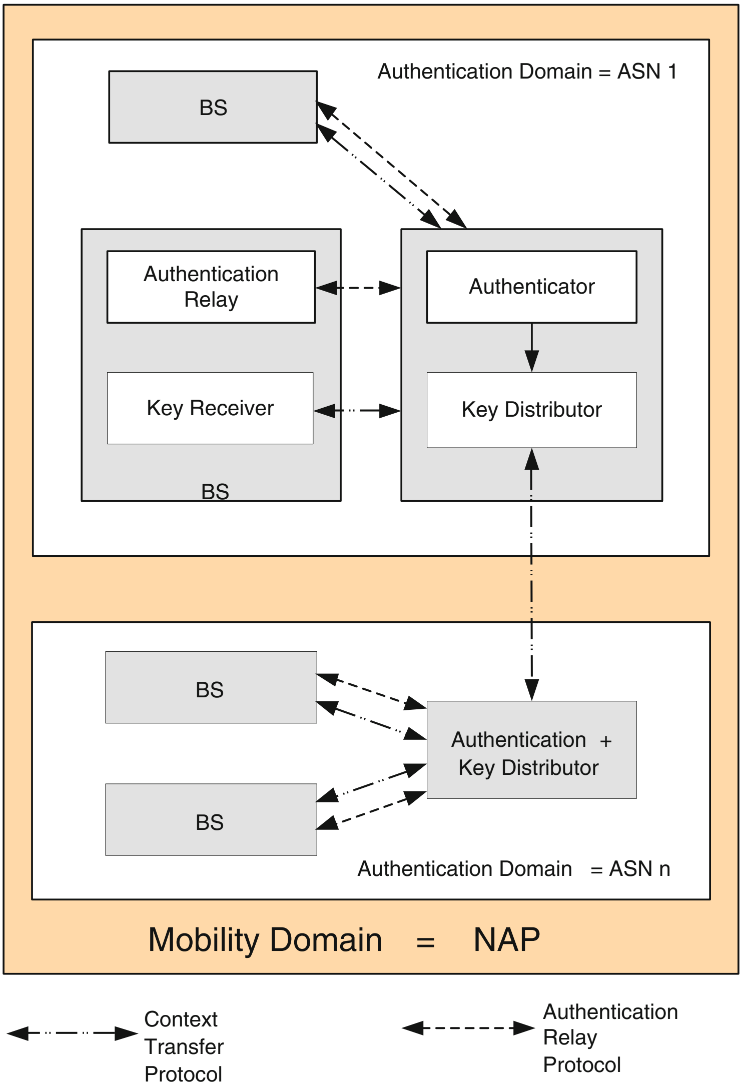

Authentication Domain = ASN 1
BS
Authentication Relay
Key Receiver
BS
Authenticator
Key Distributor
BS
BS
Authentication + Key Distributor
Authentication Domain = ASN n
Mobility Domain = NAP
Context Transfer Protocol
Authentication Relay Protocol

Fig. 10.10 Mobility and authenticator domains – Standalone model (© WiMAX Forum 2005–2007)

PKMv2 STA TEK RSP, which includes the SA descriptor list. For each SA, two Traffic Encryption Keys (TEK) are derived by the BS and sent to the MS. These are encrypted with KEK (Key Encryption Key) as the symmetric secret key. After three-way handshake, registration procedure starts. Completion of registration procedure initiates service flow creation.

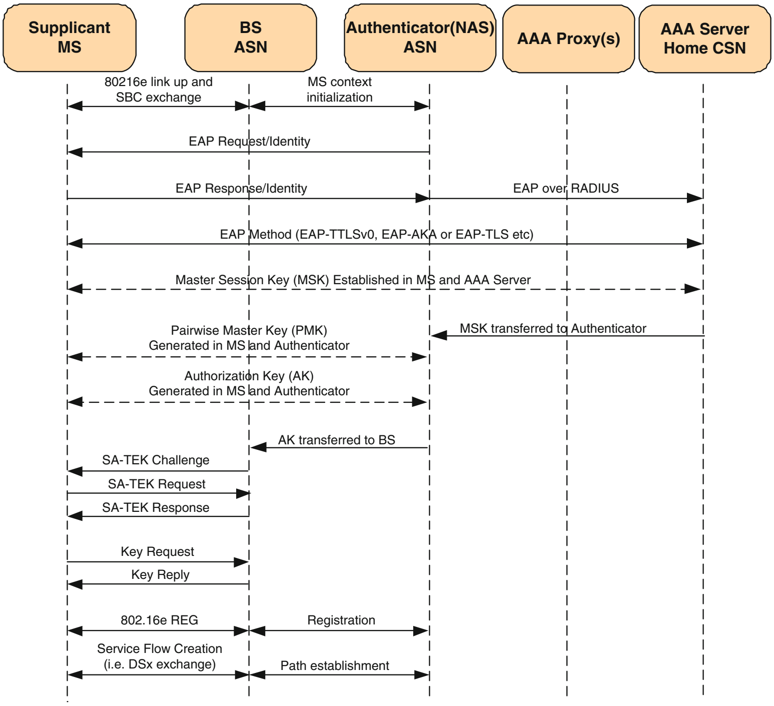

Supplicant MS
BS ASN
Authenticator(NAS) ASN
AAA Proxy(s)
AAA Server Home CSN
80216e link up and SBC exchange
MS context initialization
EAP Request/Identity
EAP Response/Identity
EAP over RADIUS
EAP Method (EAP-TTLSv0, EAP-AKA or EAP-TLS etc)
Master Session Key (MSK) Established in MS and AAA Server
Pairwise Master Key (PMK) Generated in MS and Authenticator
MSK transferred to Authenticator
Authorization Key (AK) Generated in MS and Authenticator
AK transferred to BS
SA-TEK Challenge
SA-TEK Request
SA-TEK Response
Key Request
Key Reply
802.16e REG
Registration
Service Flow Creation (i.e. DSx exchange)
Path establishment

Fig. 10.11 PKMv2 procedure during network entry

10.10.2 Authenticator and Mobility Domains

Two concepts of operation are specified: authenticator domain and mobility domain. Authenticator domain consists of one or more BSs, which are serviced by a single authenticator but a BS may belong to more than one authenticator domain. During network entry of MS, BS forwards the EAP messages to an authenticator, which becomes MS's anchor authenticator. On the other hand, mobility domain may consist of more than one authenticator domain in which a single PMK is used to derive the keys and context is transferred during the handover. Mobility domain may be equal to NAP but note that PMK sharing is not allowed. Figure 10.10 depicts the relation between authenticator domain and mobility domain for standalone mode. During handover, context transfer happens between key distributors.

10.11 Accounting

The accounting framework in WiMAX considers RADIUS-based accounting for NWG Release 1. Accounting architecture is shown in Fig. 10.13 in which accounting agent in ASN makes a connection with home AAA server through visited AAA

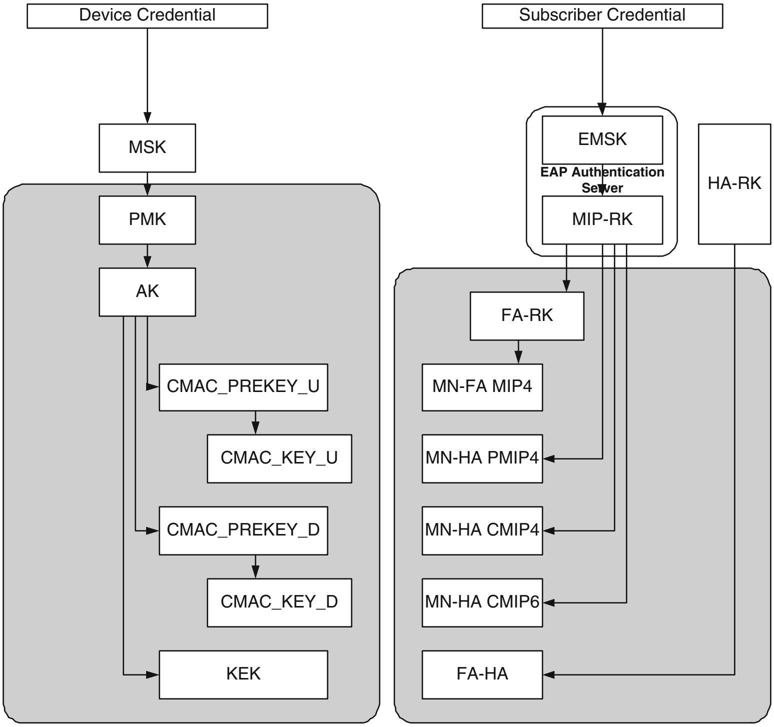


Fig. 10.12 Key hierarchy

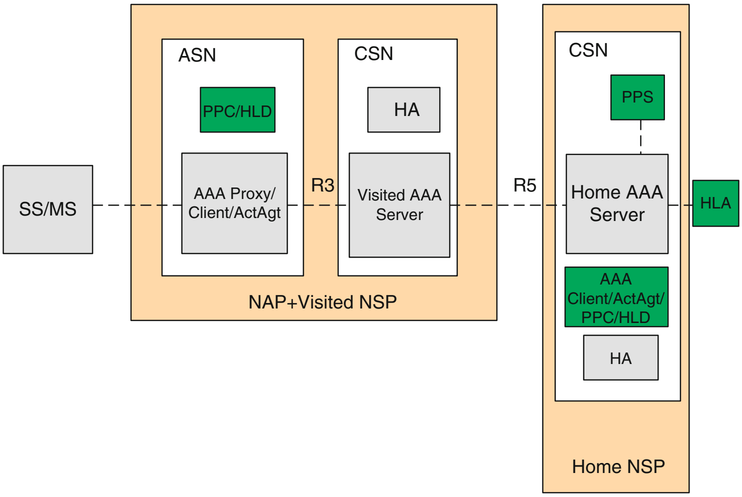


Fig. 10.13 Accounting architecture; Discarded or unsent data between MS and the account agent cause inaccurate charging; the accounting agent informs the Negative Volume count to AAA to avoid overcharging

server over RADIUS protocol. There are three types of accounting methods: offline (postpaid), online (prepaid), and hot-lining support.

10.11.1 Offline Accounting

In offline accounting, packet data accounting has a tiered architecture: Airlink and IPlink records. Airlink records are mainly number of bytes/packets dropped at the BS. The serving ASN merges Airlink and IPlink records into User Data Records (UDRs) and sends to home AAA server.

UDR is created when the R6 connection is established. UDR can be transferred when there is handover. ASN uses *RADIUS-Accounting-Request-Start* and *RADIUS-Accounting-Request-Stop* messages to start and stop the accounting in the server.

10.11.2 Online Accounting

Online accounting on the other hand introduces a packet data service with fixed volume and duration in advance. Account status is stored in a prepaid server residing in the home CSN in connection with Home AAA server.

Prepaid client (PPC) that can reside either in ASN or CSN is responsible to track the traffic per user. Tariff switching is also possible to change the tariff for different time of a day. Account balance is updated by the Home AAA or Prepaid Server (PPS) according to the consumption. *RADIUS-Access-Accept* message coming from the server is relayed to the PPC to inform the user about the remaining balance.

10.11.3 Hot-Lining

Hot-lining is used to prevent blocking data access arbitrarily. It is an interruption of the data service at the start of packet data session or midsession in order to direct the data service to a Hot-Line Application (HLA) to notify the reason. Reasons include depleted prepaid account, billing issues, expiration of a credit card, etc. This is to notify user to prevent blocking data access arbitrarily (Fig. 10.14).

There are two methods to hot-line a user: profile-based hot-lining or rule-based hot-lining. In profile-based hot-lining, Home AAA sends a hot-line profile identifier in the RADIUS message, and hot-line profile identifier selects a set of rules in the hot-line MS (HLD) to determine whether user's session is redirected or not. In rule-based hot-lining, home AAA sends the actual redirection rules and filter rules to redirect and/or block the user's sessions.

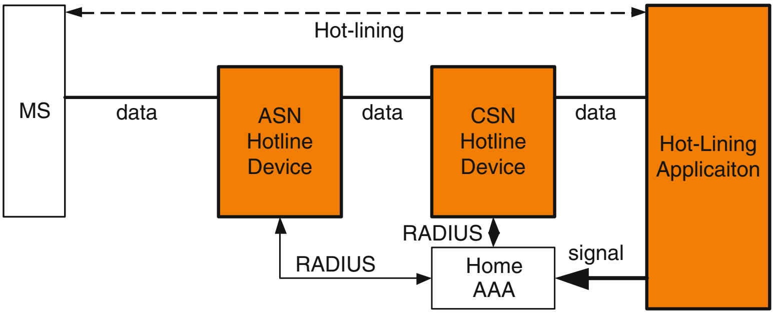


Fig. 10.14 Hot-lining

In addition to these, QoS-based accounting is also supported where each data flow is differentiated and certain accounting methods are applied depending on the characteristic of a flow.

10.12 QoS framework

The QoS framework[6] defined in the WiMAX Forum complements the QoS framework of IEEE 802.16. QoS framework comprises the following elements:

- Connection-oriented service
- Delivery services: UGS, RT-VR, ERT-VR, NRT-VR
- Provisioned QoS parameters per subscriber
- Policy-based admission
- Static or dynamic service flow creation where the latter can create, modify, or delete service flows dynamically.

Extension of IEEE 802.16 QoS structure to core networking requires additional functional elements such as Policy Function (PF), Admission control (AC), and Service Flow Authorization (SFA). Figure 10.15 specifies these elements with respect to their hosts. The PF resides in the NSP along with the policy database. The AAA works in conjunction with the PF to consider user QoS credentials and associated policy rules. The SFA, residing most likely in the ASN-GW, uses this information during network entry and communicates to the service flow management (SFM) entity of the serving BS.

SFA administers ASN level policy and enforces local policy database, which is associated to local policy function. There is an anchor SFA for each MS, which stays permanent during the time of device authentication. Anchor SFA can uses relay

[6] The WiMAX forum notes that this could be implementation specific; the NWG releases make no guarantees.

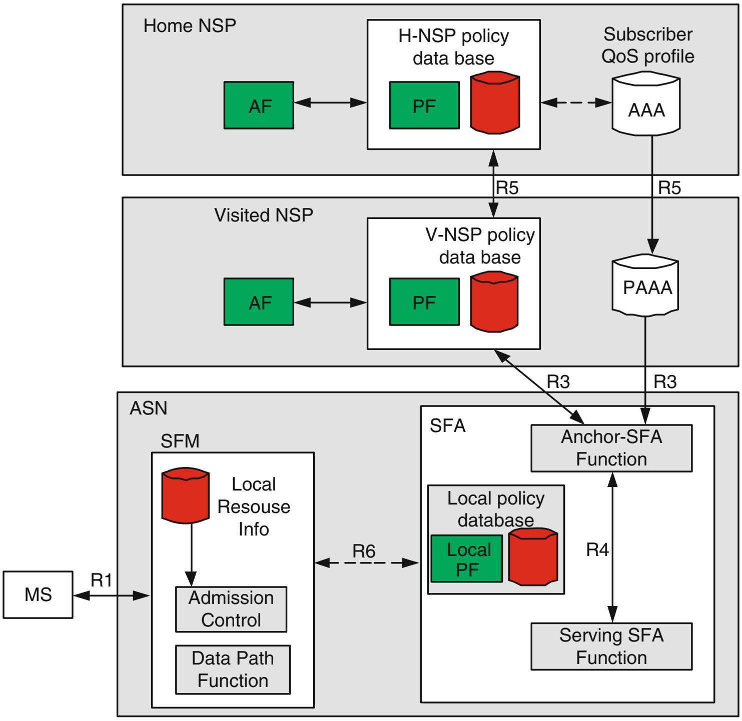


Fig. 10.15 QoS architecture (© WiMAX Forum 2005–2007)

SFAs to communicate to the SFM where relay SFA that communicates directly with the SFM is called *serving SFA*. Anchor SFA keeps track of the current serving SFA when MS moves in the network.

SFM creates and grants admission, modifies and deletes service flows over the airlink. Admission Control function is responsible to have the final decision for the service flow request coming from the SFM since it monitors local and radio resource usage and decides accordingly.

When the QoS profiles are downloaded, the SFA creates, admits, and activates the Pre-Provisioned Service Flows. Also, if the user's QoS profile has not been downloaded, then the PF can initiate the creation of preprovisioned service flows as seen in Fig. 10.16.

10.12.1 DiffServ Support

Differentiated services (DiffServ) is a well-known IP layer QoS mechanism in which network point of entry and network elements assign *code points* to enforce priority of packets. The type of traffic to be serviced with different priority is selected by the network operator since DiffServ only provides a framework to allow classification and differentiated treatment.

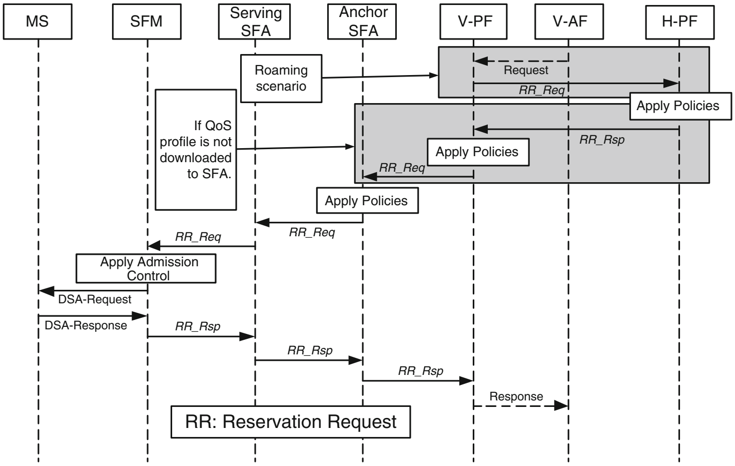


Fig. 10.16 Service flow creation triggered by the AF at the visited NSP

In WiMAX air link, DiffServ can be used in two ways: first, it can be used within a service flow to enforce priorities for packets by injecting code points in ASN or MS since multiple type of traffic can be carried over a single flow. For example, if there is one RT-VR service flow, voice packets within a flow can be of high priority and the rest can be of low priority. Second, it can be used to establish service flows based on DiffServ classes. This happens if packets are already marked with a service flow before entering ASN or MS. For instance, if there is a packet with high-class DiffServ code point, a UGS service flow is created automatically.

10.13 ASN Anchored Mobility

Subscriber station moves and changes the attachment point. If the attachment point (BS) belongs to the same ASN domain, MS does not need to change its IP address (PoA). This type of mobility is called *ASN Anchored Mobility*. However, there is also *CSN Anchored Mobility*, which happens if MS moves to another BS that belongs to another ASN.

Mobility management is responsible to minimize and eliminate the packet loss, reduce handoff latency and maintain packet ordering with proper security. There are three components that assist the mobility within ASN as seen in Fig. 10.17. *Handover Function* decides for handover and performs the necessary signaling to establish the communication of MS with target BS along with *Context Function*. Handover Function also initiates the *Data Path Function* along with *Data Integrity* in order to change the data path to a new location with minimum packet loss and latency.

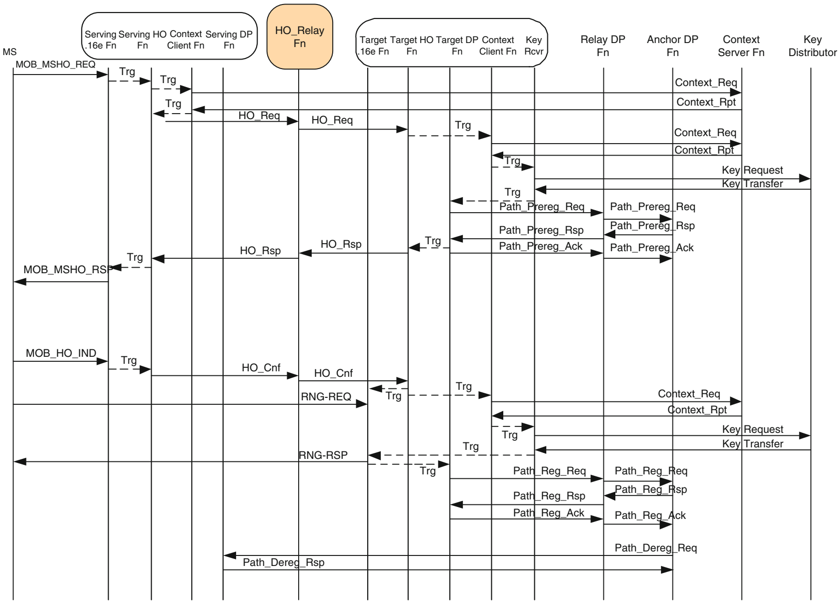


Fig. 10.17 ASN mobility functions: if BS can communicate through R8 link, then Relay HO Function is not used during R8 Handover (© WiMAX Forum 2005–2007)

10.13.1 Data Path (Bearer) Function

The data path function is responsible to set up bearer plane between two peers with tunnels. Bearer plane can be between BSs or between ASN-GWs or between ASN-GWs and BSs. The data path function may be classified by its roles: anchor, serving, target, or relay DP function. Anchor DP function is the anchor point for data and forwards the packets toward serving DP function that resides in the BS, which has connection to particular MS through air link. During handover, anchor DP function replaces the data path toward target DP function residing in the new BS and uses relay DP functions if necessary to reach serving DP or target DP.

There are two types of data path bearer: Type-1 and Type-2. Type-1 is a generic Layer 3 tunnel (e.g., IP-in-IP or GRE) and Layer 2 (Ethernet or MPLS) network. The payload is IP datagram or Ethernet packet. Each Type-1 bearer is classified according to a SFID, which is sent once during Layer 2 setup and is not modified during HOs. More coarser classification is also possible such as creating a Type-1 bearer per MS, or per BS. According to Type-1 bearer, packets are classified in the anchor DP function (most likely resides in ASN-GW) and sent to serving DP function with specific key over a downlink data path tunnel. Serving DP function is responsible to transmit packet to MS with correct CID, which is refreshed whenever MS moves to a new BS. Also, MS classifies the uplink packets based on CID.

Type-2 is same as Type-1 but now the data packet is defined as an 802.16e MAC Service Data Unit (SDU) with additional information such as CID of target BS, ARQ Parameters. Type-2 carries Layer-2 data packets as opposed to Type-1. Anchor DP function can reside in a BS, which may have a direct 802.16 physical association with MS.

10.13.2 Handoff Function

Handoff function is responsible for the HO decision as well as operation and signaling procedures related to HO. There are several handover types supported: mobile-initiated, network-initiated, and fast handover such as FBSS and MDHO. Mobile-initiated handover and network-initiated handover differ only in the initiation of the handover. Either MS decides for handover with respect to its signal readings or base station (or ASN-GW in Profile A) decides for handover with respect to its populated signal readings and capacity information.

Handover function in the serving BS controls the handover procedure and communicates with its peer, handover function of target BS. Necessary context residing in the MS or in the network is consolidated through the context server and client.

10.13.3 Context Function

MS-related context in the network needs to be transferred or updated to new BS. Context in the serving/anchor handover function needs to be updated and transferred to target HO function. Most updated information is kept in the Context Server and exchanged with Context Client, residing in the BS that has a direct connection to MS through air link.

10.13.4 Data Integrity

During handover, minimal disruption is aimed to reduce the impact of HO. Maintaining data integrity during HO reduces the rate of packet loss, duplication or reordering together with reducing the datapath setup latency/jitter. There are two available mechanisms for data integrity, which are mutually exclusive and can be combined for better performance: buffering and bi/multicasting.

- *Buffering*: Any data path function can buffer the packets in order to be sent later through the air interface.
- *Bi/Multi-Casting*: Anchor Data Path function or serving data path function can bi/multicast the downstream traffic to the target BSs through bi/multiple data paths.

These packets are consolidated in the target BS and the packets that have not been transmitted are sent to MS after handover. If only buffering is performed in the serving DP function, the serving DP function may forward the buffered and not transmitted packets to target BS over R6 or R8 link, if available, and/or serving data path function may create a temporary data-path to target data path function until anchor data path switches the data path to target data path function.

10.14 CSN Anchored Mobility

When MS moves to another BS that belongs to another ASN mobile IP-based mobility takes place, which requires reanchoring the current Foreign Agent (FA) to a new FA as seen in Fig. 10.18. CSN Anchored Mobility administers the establishment between ASN and CSN that are in or different domains.

CSN Anchored Mobility in IPv4 network can be performed by two different methods: Proxy MIP and Client MIP, which has higher priority if they coexist. In Proxy MIP, ASN performs the role of the MS and communicates with the HA to do the MIP registration. As a result, MS is agnostic to the CSN mobility management activities. In CMIP procedure, MS by itself initiates MIP procedures.

10.14.1 Proxy MIP

PMIP Client resides in ASN-GW with Authenticator and establishes MIP signaling between FA and HA, which is assigned by home or visited NSP as seen in Fig. 10.19. MS does not need any additional requirements to support CSN Anchored Mobility. The necessary information such as HA address, security information, DHCP server address, etc are retrieved from AAA messages during authentication.

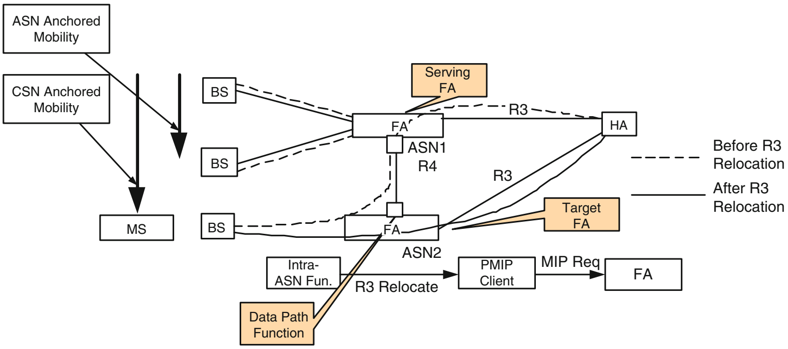


Fig. 10.18 CSN Mobility (© WiMAX Forum 2005–2007)

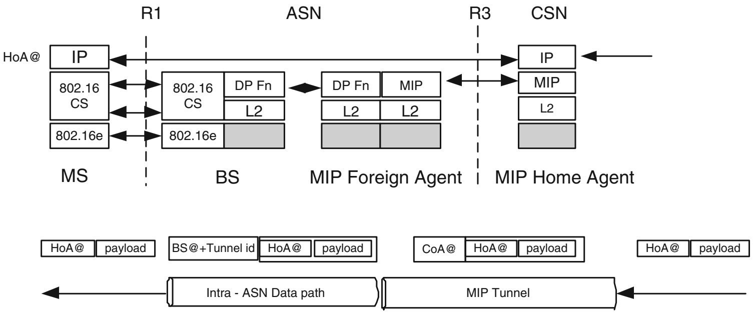


Fig. 10.19 Proxy MIP example (© WiMAX Forum 2005–2007)

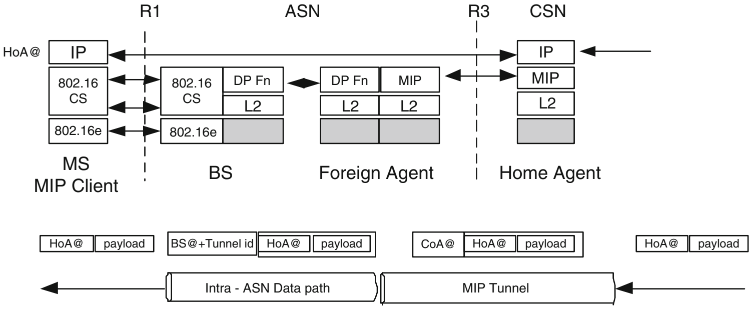


Fig. 10.20 Client MIP example (© WiMAX Forum 2005–2007)

There can be two different types of DHCP deployment: if there is a DHCP server, DHCP relay can be located in the ASN to relay the DHCP message to the server, or DHCP proxy in ASN (resides in ASN-GW) can manage the DHCP exchange with MS, which does not require a DHCP server in the CSN.

IP Address (HoA address) can be assigned either by DHCP server, AAA, or Home Agent. In the AAA scheme, HoA is available during authentication. In DHCP server method, IP address is retrieved during the DHCP exchange, and in MIP method, DHCP proxy waits till the MIP registration completes and sends the HoA with DHCP offer message. DHCP renewals are initiated by the MS and processed either by DHCP relay/server or DHCP proxy.

10.14.2 Client MIP

In Client MIP, MIP Client resides in the MS above the 802.16 drivers as seen in Fig. 10.20. MIP Client participates in the message exchanges required to perform CSN Anchored Mobility. After successful authentication, FA advertisements are sent to MS. Then, MIP registration is initiated by the MS with HA via FA. HA communicates with the Home AAA, and IP host configuration can be relayed back in the MIP Registration Reply Message.

CMIPv6 for IPv6 is similar to CMIPv4 in many ways except that there is no FA. MS uses a colocated care-of address (CoA) to communicate with the HA to do the binding update and additionally, correspondent host can directly communicate with the MS bypassing the HA with route optimization. MS can retrieve a CoA from DHCPv6.

10.15 RRM: Radio Resource Management

RRM is responsible to utilize the radio resources efficiently. RRM function collects information about wireless link capability or available spare capacity and may assist other modules based on these aggregated information.

RRM function comprises a Radio Resource Agent (RRA) in BS and Radio Resource Controller (RRC) in BS, ASN-GW, or a standalone server. RRA is responsible to collect radio resource information of air link and relay them to the RRC. Information in RRC can be used for several purposes:

- Admission control and connection
- Service flow admission
- Load control
- Handover decision and preparation

RRM can work in conjunction with the *Network Resource Management Entity*, which monitors the resources in the backhaul.

10.16 Paging and Idle Mode

MS enters *idle* mode to conserve energy. During idle mode, if MS is needed to be located by the network due to a data inquiry, paging operation is used. Paging operation comprises Paging Controller (PC), Paging Agent (PA), Paging Group (PG), and Location Register (LR).

Paging Controller is responsible to administer the activity of MS in idle mode. *Paging Agent* residing in BS communicates with the PC and applies 802.16e paging operation (see Chap. 9). A BS is part of one or more paging groups, which are set by the *Network Management System*. *Location Register* is a database to maintain the location of MS during idle mode. Paging information may include paging group, last reported BS, and service flow information, etc.

MS enters idle mode with DREG_REQ message, which initiates PA in the BS to send the request to PC. The PC verifies that the MS is allowed to enter idle mode, and may transfer some security context to Anchor Authenticator to retain some security context. The PC talks to Anchor DP function in order to release the data paths to stop data forwarding. The PC also talks to LR to create an entry for the MS. During idle mode, FA may or may not migrate to new location of MS.

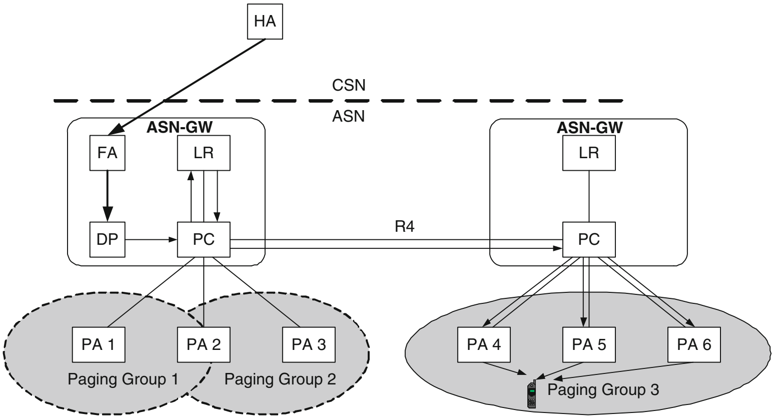


Fig. 10.21 Paging operation

MS monitors the PG during the idle mode and when MS crosses boundary of its current PG, it has to perform *Location Update* procedure in order to update the LR entry. MS can perform secure or nonsecure location update depending on the valid security context in the BS. Nonsecure location update results in MS network reentry and reauthentication as in regular network entry.

When data arrives to the data path function, MS is paged that is an announcement to PAs from the PC. All the BSs that belong to the same *Paging Group* of the MS broadcast a *paging message*. Upon receiving the paging message, MS initiates idle mode exit by sending a *Ranging Request* in order to initiate network entry (Fig. 10.21).

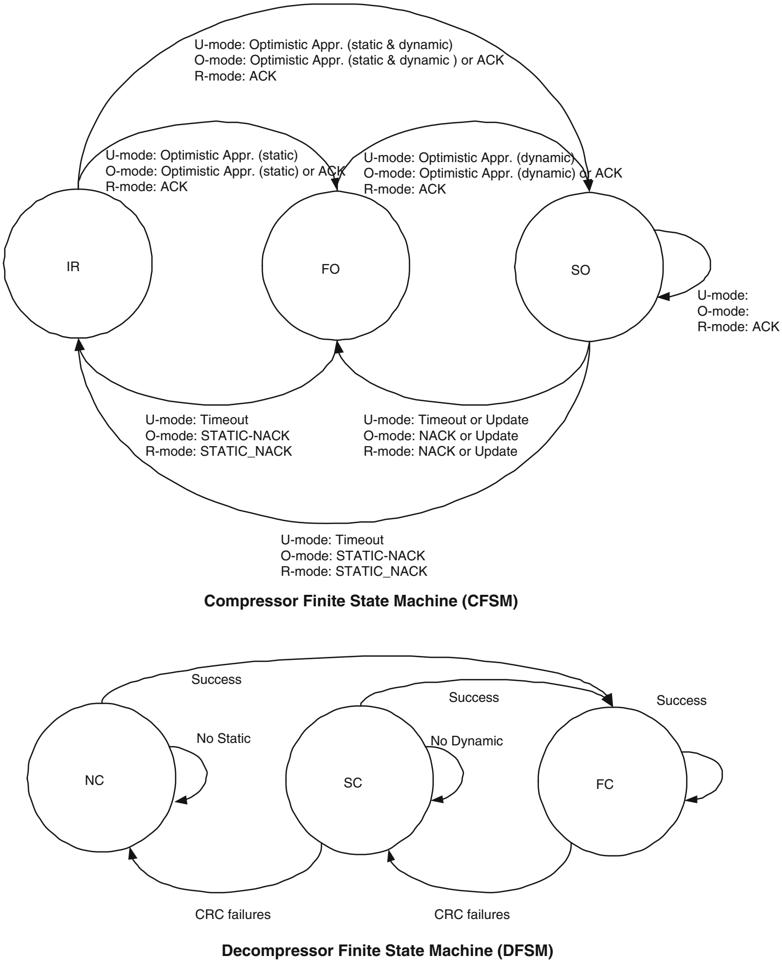
10.17 Release 1.5 Features

WiMAX technology is evolving with new features to not only facilitate easy network integration for service providers but also offer subscribers with an enhanced experience donated with next generation services. In this section, we present some of the features that are considered within WiMAX Forum NWG Release 1.5, planned to be ratified in 2008. Although much care is taken to present up-to-date information, the following features are subject to change.

10.17.1 ROHC: RObust Header Compression

The Header compression (see Sect. 3.8) compresses the protocol headers due to redundancy in the header fields of the same packet as well as consecutive packets of the same packet flow. A packet classifier residing in a data path function can identify a flow by set of some parameters such as protocol headers, the source and destination addresses, and the source and destination ports, etc. These constitute context and can be established on both sides of the link in the beginning with a few packets that are sent uncompressed. This information is used by the compressor to compress and by the decompressor to decompress the packets to its original state.

The ROHC [RFC3095] algorithm establishes a common context at the compressor and decompressor by transmitting full header and then gradually transition to higher level of compression. ROHC is designed to be flexible to support several protocol stacks and each protocol stack defines a "profile" within the ROHC framework. ROHC compression is essentially the interaction between compressor and decompressor. A ROHC compressor, seen in Fig. 10.22, is in one of three main states:


U-mode: Optimistic Appr. (static & dynamic)
O-mode: Optimistic Appr. (static & dynamic) or ACK
R-mode: ACK
U-mode: Optimistic Appr. (static)
O-mode: Optimistic Appr. (static) or ACK
R-mode: ACK
U-mode: Optimistic Appr. (dynamic)
O-mode: Optimistic Appr. (dynamic) or ACK
R-mode: ACK
IR
FO
SO
U-mode:
O-mode:
R-mode: ACK
U-mode: Timeout
O-mode: STATIC-NACK
R-mode: STATIC_NACK
U-mode: Timeout or Update
O-mode: NACK or Update
R-mode: NACK or Update
U-mode: Timeout
O-mode: STATIC-NACK
R-mode: STATIC_NACK

Compressor Finite State Machine (CFSM)

Decompressor Finite State Machine (DFSM)

Fig. 10.22 ROHC compressor and decompressor finite state machines

IR: In Initialization and Refresh (IR) state, the compressor has just been created or reset, and full packet headers are sent.

FO: In First-Order (FO) state, the compressor has detected and stored the static fields (such as IP addresses and port numbers) on both sides of the connection. The compressor is also sending dynamic packet field differences in FO state. Thus, FO state is essentially static and pseudodynamic compression.

SO: In Second-Order (SO) state, the compressor is suppressing all dynamic fields such as RTP sequence numbers, and sending only a logical sequence number and partial checksum to cause the other side to predictively generate and verify the headers of the next expected packet. In general, FO state compresses all static fields and most dynamic fields. SO state is compressing all dynamic fields predictively using a sequence number and checksum.

The states of decompressor, seen in Fig. 10.22, are No Context (*NC*), Static Context (*SC*), and Full Context (*FC*). ROHC modes described later determine the state transitions:

U-Mode: In Unidirectional mode, packets are sent in one direction, from the compressor to the decompressor. In cases where the return path or the reverse channels are not available it requires periodic refresh.

O-Mode: In Bidirectional Optimistic mode, a feedback channel is utilized. It does not require periodic refresh.

R-Mode: In Bidirectional Reliable mode, it issues feedback for all context updates.

Also, all kind of mode changes from U to O/R and O to U/R and R to U/O are possible at any time.

The ROHC employs Window-based Least Significant Bits Encoding (W-LSB), which is similar to video compression, in that a base frame and then several difference frames are sent to represent an IP packet flow. This has the advantage of allowing ROHC to survive many packet losses in its highest compression state, as long as the base frames are not lost. An example is depicted in Fig. 10.23.

W-LSB is a variant of LSB in which *i* significant bits of the field are transmitted and decompressor uses the original value with the reference value *r* that includes the untransmitted bits. W-LSB brings robustness to the LSB when the compressor is unable to determine the exact reference value. The compressor maintains a sliding window of possible reference values and chooses the minimum number of least significant bits (*i*) that will produce the original value given those reference values.

ROHC is more powerful over PHS since PHS cannot suppress the field "Sequence number" and "Time stamp," of which the second-order difference is 0 since the first-order difference is constant as seen in Fig. 10.24.

In addition, PHS cannot suppress "Payload type" even though that field is static, because PHS operates as the unit of byte, and the first bit of the second byte (Marker bit) is not static to suppress. ROHC compresses RTP header to 2 bytes when the second-order differences of the fields are all 0.

PHS uses PHSM (Marker) to identify whether the marked byte shall be suppressed or transmitted. Therefore, PHS works only for the case when the first-order

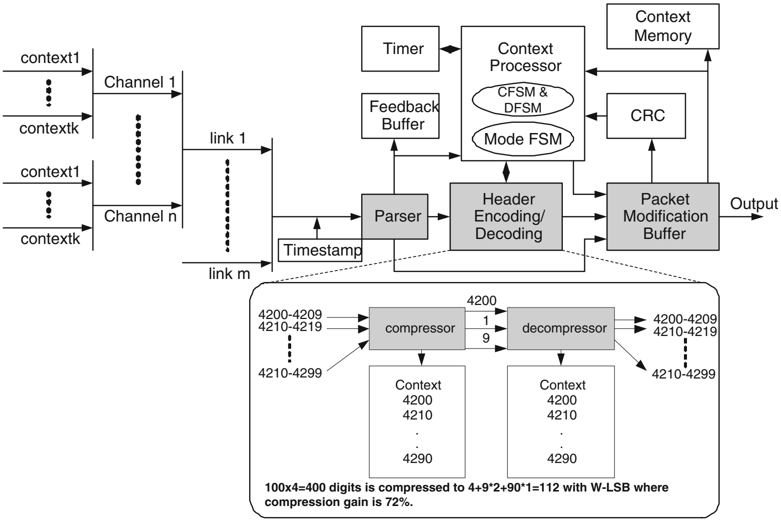


Fig. 10.23 ROHC framework

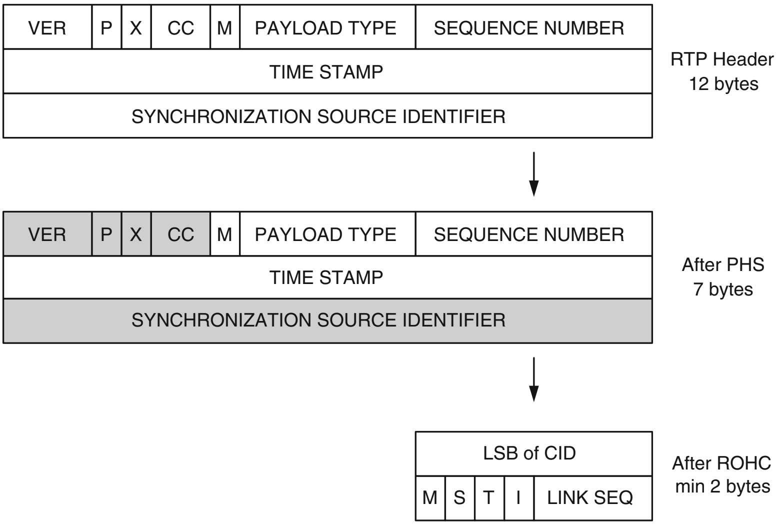


Fig. 10.24 ROHC vs. PHC

difference between the previous packet and the current packet is 0. ROHC compresses the fields when not only the first-order difference is 0, but the second-order difference is 0. Even though the second-order difference is not static, it compresses the fields by using of delta encoding.

ROHC support in WiMAX requires signaling between ROHC and CS in 802.16e MAC. ROHC function in ASN (may reside in ASN-GW) and SFA collocated with

Anchor Data Path function perform header compression and decompression. In ASN-GW, SFA may receive ROHC policy from either AAA or PCRF (see last chapter) and generates the classification rule for ROHC. SFA exchanges the information about service flows with ROHC of MS via SFM in BS. If MS accepts, then SFA initiates ROHC parameter negotiation. ROHC function initiates per-channel negotiations, which is done through DSx-REQ/RSP MAC messages. After negotiation, DP function sends the packets to ROHC compression if DL packet belongs to a ROHC channel. In MS, a packet in ROHC channel is sent to ROHC de-compressor. For uplink, ROHC compression is performed in MS and de-compression is performed in DP function. During the idle mode, ROHC context can be stored in the Paging Controller.

10.17.2 MCBCS: Multicast Broadcast Services

The Multicast Broadcast Services within WiMAX network complement the MBS framework specified in IEEE 802.16e-2005. The MCBCS usage scenarios include but are not limited to streaming of multimedia; local broadcast to notify subscribers about local events like concert, sporting game, etc.; interactive online gaming; push-to-talk; file downloading; prescheduled downloading files; downloading service guides; alert services such as traffic, weather, etc.; podcasting; interactive TV; real-time monitoring; multiparty conference call; video-on-demand, etc.

The MCBCS solution may be compatible with both Multi BS (reuse 1) and Single BS (reuse 3) deployments. The MCBCS may utilize the standard IP multicast routing protocols such as PIM-SM, DVMRP, etc. in the backbone in which packet duplication is performed by multicast routers toward ASN, CSN. There may be an MBS server in the ASN per an MBS Zone as seen in Fig. 10.25 where MBS server keeps the list of BSs belonging to the MBS zone. MBS server is responsible to connect to a multicast tree of the server via IGMP in order to receive the content. Then, it maps the content to the appropriate CID and SA in order to send to the MBS Zone group. If Multi-BS is supported then it helps BS to be synchronized in frame and burst. MCBCS stops sending the packets if there is no MS in the MBS Zone requesting a MCBCS service. MBS Server may also communicate to SFA and Paging Controller to initiate or terminate the MCBCS services.

10.17.3 LBS: Location Based Services

Location-aware application capability is an important feature for next generation networks. WiMAX network considers location capability to enable Location-Based Services, Emergency calls, Lawful intercept, and features to optimize the network such as location-assisted handover or load balancing. LBS network reference model

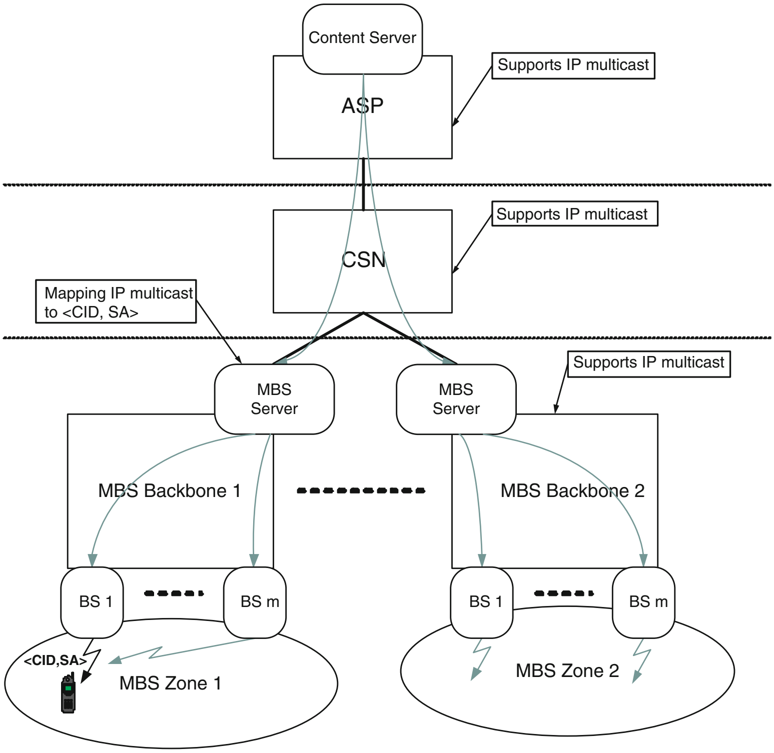


Fig. 10.25 End-to-end MCBCS scenario

consists of Location Server (LS), Location Controller (LC), and Location Agent (LA) as seen in Fig. 10.26.

LS: The Location Server, located in CSN, is the central point that queries location information from the access network as well as provides the location information to authorized entities. For a device to activate LBS, LS communicates to AAA to get the Authenticator ID of the device and retrieves the *Location Controller ID* from the authenticator.

LC: The Location Controller, residing in ASN-GW, is responsible to get the location information of the device through either device-based or network-based methods or mix of both. LC knows the serving BS of the device and is responsible to trigger the signaling required for measurement of location.

LA: The Location Agent, located in BS or MS, is responsible to make and report measurements.

There can be two types of services: MS-managed location or network-managed location. MS-managed location service utilizes the LBS feature as a navigation service with or without a GPS capability. MS subscribes to this service during network

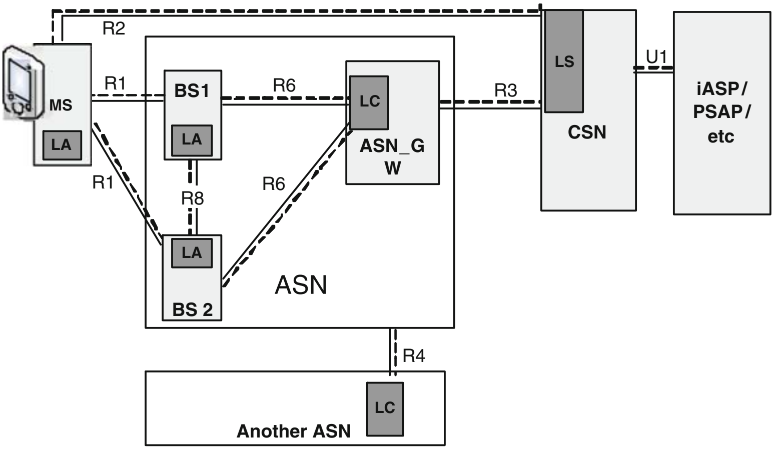


Fig. 10.26 Reference model for Location-Based Services (© WiMAX Forum 2005–2007)

entry to receive geo-location information of the serving as well as neighboring BSs via LBS-ADV message. MS uses this information and relays it to its applications.

Network-managed location can be initiated either by mobile or by network itself. MS may have GPS, information of which may be sent to network to enhance the accuracy. WiMAX network also may broadcast *satellite ephemeris data* to decrease the time to first-fix (TTF) for the GPS to obtain the satellite fixes for measurement.

Reference BSs send reference signals and location measurements in downlink, which could be based on relative delay, round trip delay, or RSSI and in uplink it could be based on UL sounding or the RNG-REG as seen in Fig. 10.27.

MS-managed location information uses downlink reference signals to do triangulation; reference BSs communicate to accurately synchronize the timing. Network-managed location either only seeks serving BS geographical location of the MS with an estimated coverage with respect to CINR or cell range, set during network planning. Network-managed location may also depend on MS to do location calculation with downlink reference signals. MS then sends this back to the network. Network-managed location may also perform the location calculation in the network based on the uplink measurements.

10.17.4 ES: Emergency Services

WiMAX network is being designed to support emergency services (ES) for WiMAX commercial VoIP services. Overall WiMAX network architecture is enhanced with VoIP service provider (VSP) and ES provider network infrastructure (PSAP[7]) as seen in Fig. 10.28. Two reference points introduces are R_E and R_V:

[7] Public Safety Answering Point in US, Operator Assistance Centres in UK, and Communications Centres in New Zealand.

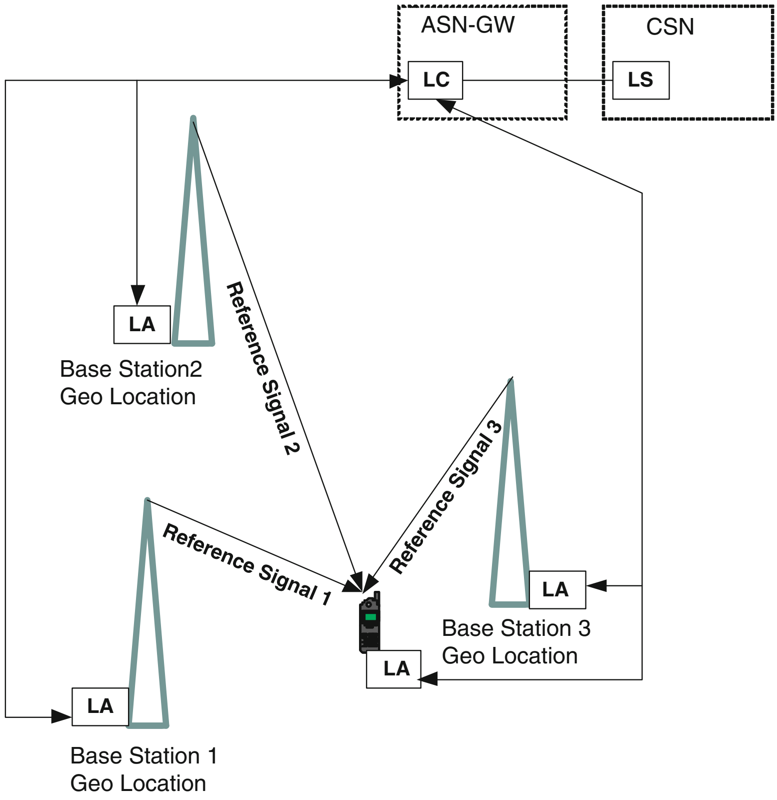


Fig. 10.27 Location determination with reference signals

R_E: Reference point for Emergency Services defines the protocols and procedures between VoIP service and PSAP to handle emergency calls.
R_V: Reference point to interconnect NSP and VSP to provide VoIP support.

Supported scenarios include roaming and nonroaming case. To request for emergency services, $\{sm = 2\}$ code is used in front of the NAI as follows $\{sm = 2\} < user - name > @ < NSPRealm >$ during EAP authentication procedure. If there is no roaming then VSP attached to the NSP is used; otherwise, PSAP of visited NSP or PSAP of home NSP can be used. Also depending on the location information, a suitable PSAP can be selected.

10.17.5 LI: Lawful Intercept

The Lawful Intercept (LI) feature provides the interfaces from the network to law enforcement agencies for lawful interception as required by several national

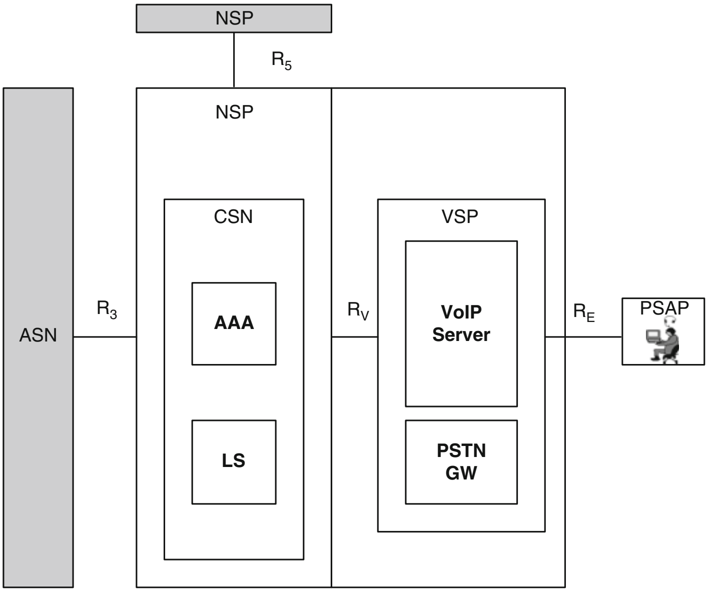


Fig. 10.28 Emergency Services (© WiMAX Forum 2005–2007)

regulations for broadband packet data. The LI operation is done in accordance with the applicable national or regional laws and technical regulations.

The LI functionality comprises *Access Function* (AF), *Delivery Function* (DF), *Collection Function* (CF), Service Provider Administration, and Law Enforcement Administration. Fig. 10.29 depicts these functions with respective owners and their interfaces to the WiMAX network.

AF consists of one or more Intercept Access Points (IAPs) that are co-located with one or more of the WiMAX ASN and CSN network elements. IAP is responsible for wiretapping based on CALEA[8] and reporting interception events/ information to the Lawful Interception Server (LIS). IAPs may be of two types: communication identifying information (CII) or communication content (CC). The IAP collects and delivers the intercepted information to the LIS. The ADMF and DF act as a mediation function to hide multiple activations by different LEAs (lawful enforcement agencies) from the IAPs on the same target. The DF delivers intercepted communication in the form of CC and CII to the CF through "E" interface. The Administration Function (ADMF) and the Delivery Function (DF) are connected through "D" interface, and every physical IAP is linked to ADMF with Li-1 interface, and to DF with Li-2 interface.

[8] "CALEA, the Communications Assistance for Law Enforcement Act, has been growing in scope since it was first passed in 1994. The initial bill was designed to make it easier for prosecutors and federal agents to tap telephones, but in 2004 the FCC extended CALEA to include broadband providers and VoIP companies..."

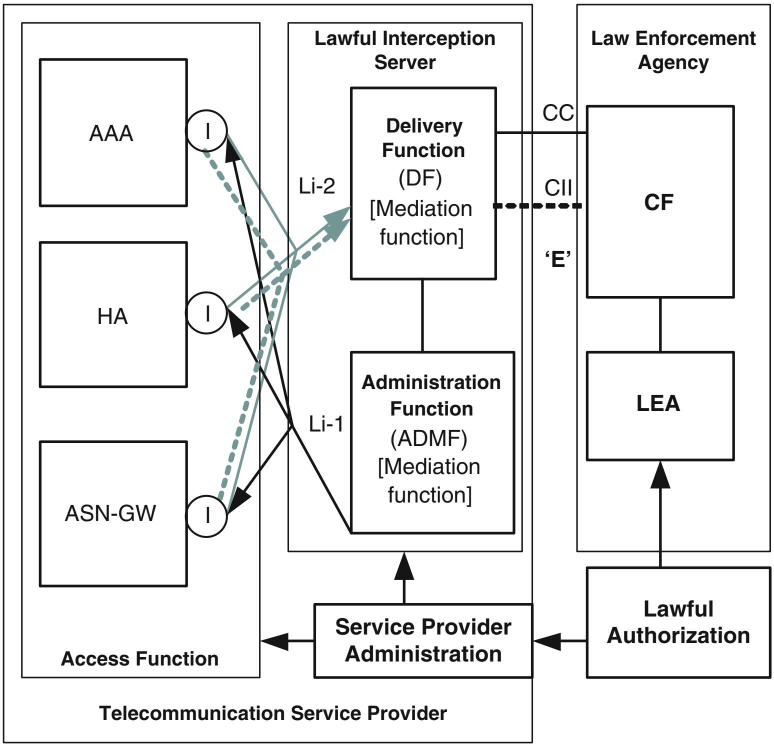


Fig. 10.29 Lawful Interception Reference Model for WiMAX: I represents IAP and TSP may represent a NAP, or NSP, or a NAP+NSP deployment case, or a "NAP Sharing" deployment case (© WiMAX Forum 2005–2007)

10.17.6 USI: Universal Services Interface

The Universal Service Interface is a framework to open WiMAX interfaces to trusted third-party Internet Application Service Providers (iASPs) (e.g., Google, E-Bay, Yahoo!). USI gives applications to do dynamic service creation as well as ability to use the intelligence being built into WiMAX network (e.g., LBS, MCBCS, PCC).

USI system resides in CSN and introduces U1 interface to ASP/iASP as seen in Fig. 10.30. *Web Services Logic* interacts with Network Capability of the network. USI framework employs access security to facilitate the requests coming from i/ASPs; the i/ASP needs to be authenticated and authorized. Usage of the network is managed with *service usage policy* that specifies some constraints to the ASP including limitations such as identity hiding. For example, the i/ASP can request the location of the MS after getting authenticated to use this service. Other services may include MS status, MS IP address discovery, etc.

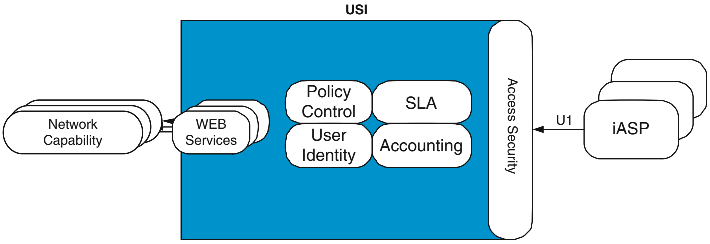


Fig. 10.30 USI module (© WiMAX Forum 2005–2007)

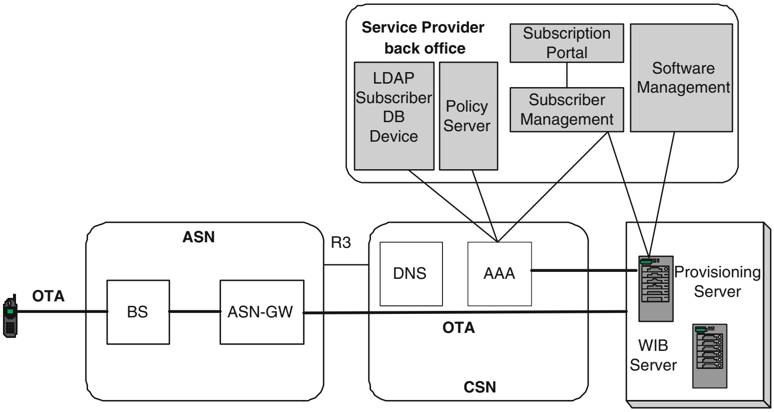


Fig. 10.31 OTA (© WiMAX Forum 2005–2007)

10.17.7 *OTA: Over-the-Air Provisioning*

Over-the-air provisioning is a feature for a WiMAX service provider to perform dynamic OTA provisioning solution in order to configure and manage the devices. OTA provides great deal of flexibility to enable open mobile devices that can be activated and enabled over the air. A device can be configured with the same device with a temporary attachment given by the service provider to initiate provisioning. Also, provisioning can be initiated by entities other than the client device. Also, a client device can initiate provisioning for other devices as well.

Figure 10.31 depicts the OTA architecture based on WiMAX network model, and Fig. 10.32 depicts the flow chart that illustrates the OTA process. *Provisioning server* and *provisioning client* are introduced in the network, which complies OMA-DM[9] solution and DSL Forum TR-069 protocol for WiMAX customer premise equipment (CPE) only. Provisioning server is responsible to perform specific management on a device via provisioning client, residing in the device.

[9] Open Mobile Alliance for Device Management (DM).

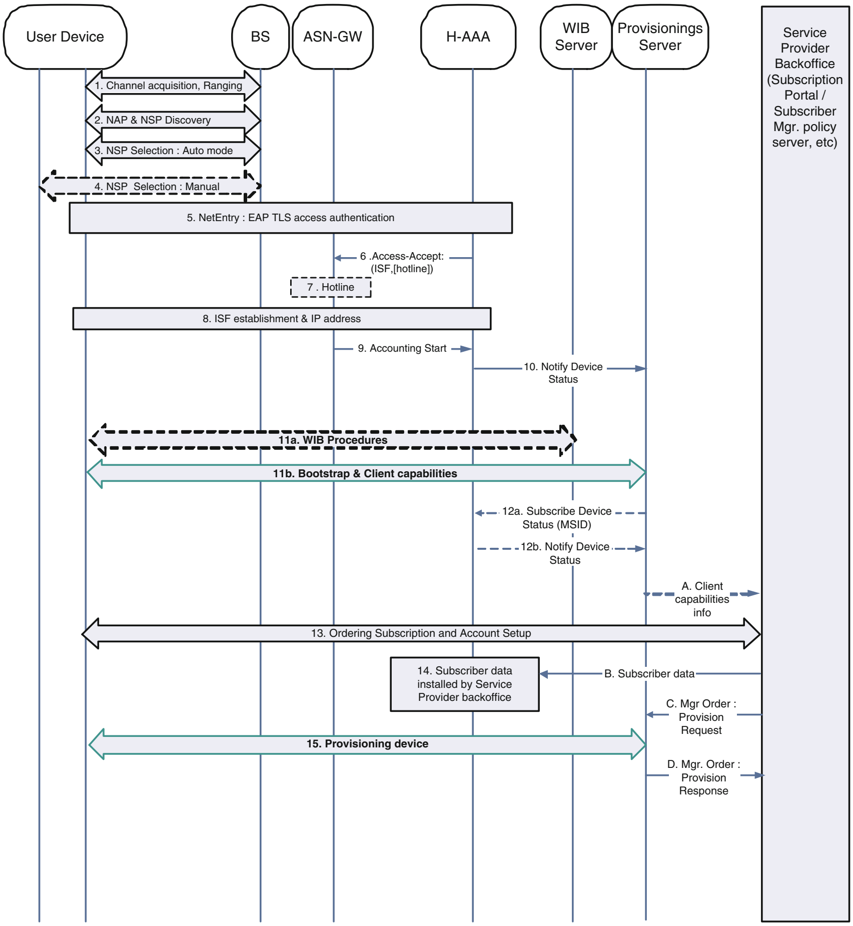

User Device
BS
ASN-GW
H-AAA
WIB Server
Provisionings Server
Service Provider Backoffice (Subscription Portal / Subscriber Mgr. policy server, etc)
1. Channel acquisition, Ranging
2. NAP & NSP Discovery
3. NSP Selection : Auto mode
4. NSP Selection : Manual
5. NetEntry : EAP TLS access authentication
6 .Access-Accept: (ISF,[hotline])
7 . Hotline
8. ISF establishment & IP address
9. Accounting Start
10. Notify Device Status
11a. WIB Procedures
11b. Bootstrap & Client capabilities
12a. Subscribe Device Status (MSID)
12b. Notify Device Status
A. Client capabilities info
13. Ordering Subscription and Account Setup
14. Subscriber data installed by Service Provider backoffice
B. Subscriber data
C. Mgr Order : Provision Request
15. Provisioning device
D. Mgr. Order : Provision Response

Fig. 10.32 OTA (© WiMAX Forum 2005–2007)

10.18 Summary

This chapter presented an overview of Mobile WiMAX network architecture. It is important to note that in this book we gave more emphasis to WiMAX network model as an IP-based network architecture when compared with other technologies that adopt the similar IP-based network. Interested reader may refer back to this chapter for more detail while traversing in LTE or UMB chapters of this book. WiMAX Network layer is demarcated into two sections with respect to functionalities and business models:

- Connectivity Service Network – Connectivity Service Network (CSN) is defined as a set of network functions that provide IP connectivity services to the WiMAX subscribers. A CSN may comprise network elements such as routers, AAA proxy/servers, user databases, interworking gateways, and mobile stations.

- Access Service Network – Access Service Network (ASN) is defined as a complete set of network functions that is needed to provide radio access to a WiMAX subscriber. An ASN comprises network elements such as one or more base stations (BS), and one or more ASN Gateways (ASN-GW). An ASN may be shared by more than one CSNs.
- Network Access Provider – Network Access Provider (NAP) is a business entity that provides WiMAX radio access infrastructure using one or more ASNs. A NAP will lease their ASNs (WiMAX equipment) to one or more NSPs.
- Network Service Provider – Network Service Provider (NSP) is a business entity that provides IP connectivity and WiMAX services to subscribers using one or more CSNs.

Mobile WiMAX architecture defines protocols and functional entities to address network entry, AAA services, QoS, mobility, radio resource management, paging, and idle mode operation. WiMAX architecture is also evolving to offer features such as RObust Header Compression, Location-Based Services, Multicast Broadcast Services, Lawful Intercept, Emergency Services, Universal Service Interface, and Over-the-Air Provisioning, etc.

WiMAX technology is no longer considered as a standalone technology for broadband wireless access; rather, it is becoming one of the key players of mobile broadband convergence with its ability to provide next-generation services as well as internetworking interfaces to 3GPP, 3GPP2, DSL, and WiFi networks. These internetworking procedures are explained in the last chapter along with WiMAX network's interferences to common IP Multimedia Subsystem (IMS) and Policy and Charging Control (PCC).

References

1. WiMAX Forum, *Recommendations and requirements for networks based on WiMAX Forum certified products*. Release 1.0, February 23, 2006.
2. WiMAX Forum, *Recommendations and requirements for networks based on WiMAX Forum certified products*. Release 1.5, April 27, 2006.
3. WiMAX Forum, *WiMAX end-to-end network systems architecture. Stage 2: Architecture tenets, reference model and reference points*. Release 1 Version 1.2, November 08, 2007. `http://www.wimaxforum.org/technology/documents.`
4. WiMAX Forum, *WiMAX end-to-end network systems architecture. Stage 3: Detailed protocols and procedures*. Release 1 Version 1.2, November 08, 2007. `http://www.wimaxforum.org/technology/documents.`
5. *IEEE 802.16-2004 October 2004, Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, August 2004.
6. *IEEE 802.16e-2005 March 2006, Physical and Medium Access Control layers for Combined Fixed and Mobile Operation in Licensed Bands*, March 2006.
7. Bormann, C., et al, "RObust Header Compression (ROHC): Framework and four profiles: RTP, UDP, ESP, and uncompressed," *RFC3095*, July 2001.
8. Jonsson, L-E., "RObust Header Compression (ROHC): Terminology and Channel Mapping Examples," *RFC3759*, April 2004.

Chapter 11
Long-Term Evolution of 3GPP

The third Generation Partnership Project (3GPP)[1] addresses next generation IP-based OFDMA technology with Long-Term Evolution (LTE) project in order to accommodate increasing mobile data usage and new multimedia applications. This new OFDMA-based air interface is utilized under Evolved UMTS Terrestrial Radio Access Network (E-UTRAN) and supported with a new flatter-IP core. This entire architecture is termed as 3GPP EPS (Evolved Packet System); previously network architecture was called System Architecture Evolution (SAE). Note that LTE air interface is a successor of GSM/EDGE and UMTS/HSxPA network technologies but has the ability to support bandwidths wider than 5 MHz, and EPS network architecture considers reduction of network level elements when compared with HSPA (High Speed Packet Access).

LTE has set aggressive performance requirements and enhances its IP-based OFDMA access with MIMO and smart antennas. Although the specification has not been finalized yet, significant details are emerging.

LTE is being designed to address higher throughput, increased base station capacity, reduced latency, and full mobility. UTRA & UTRAN Long-Term Evolution has started in December 2004 and first commercial deployments are expected in 2010.[2] *3GPP has aproved the functional freeze of LTE as part of Release 8 in December 2008.*[3]

LTE's objectives include a radio-interface physical layer to support transmission bandwidth up to 20 MHz together with new transmission schemes and advanced multiantenna technologies. LTE performance requirements are presented in Table 11.1. LTE's perspective for mobility considers optimum performance for mobile speeds 0–15 km/h. LTE is being designed to ensure high performance between 15 and 120 km/h and still expected to maintain mobility at speeds between 120

[1] 3GPP^{TM} TSs and TRs are the property of ARIB, ATIS, ETSI, CCSA, TTA, and TTC who jointly own the copyright in them. They are subject to further modifications and are therefore provided to you "as is" for information purposes only. Further use is strictly prohibited.

[2] "In US, Verizon Wireless plans to launch its LTE network in the 700 MHz spectrum ($9.36 billion worth) around 2010 timeframe."

[3] All documentation can be found in `http://www.3gpp.org/ftp/Specs`

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_11, © Springer Science+Business Media LLC 2009

Table 11.1 LTE performance metrics

Peak data rate	DL/UL: 100/50 Mbps for 20 MHz
Full mobility	Up to 500 km/h
Latency in control/user plane	<100 ms (idle to active)/<5 ms
Capacity	>200 users per cell (5 MHz)
Cell sizes	5–100 km
Spectrum	1.25, 2.5, 5. 10,15, and 20 MHz

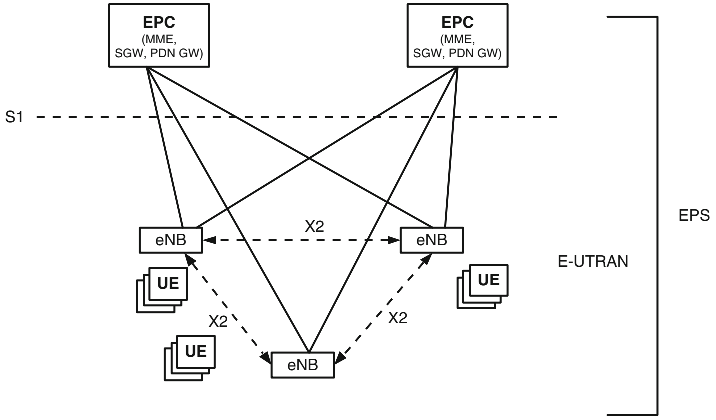


Fig. 11.1 LTE architecture

and 350 km/h even up to 500 km/h for some frequency bands. At over entire speed range, LTE is expected to support voice and real-time service quality without interruption. As one can see, mobile speeds above 350 km/h are mainly for trains. Typical performance criterion is uninterrupted operation below 120 km/h for vehicular and pedestrian speeds.

EPS framework comprises Evolved Packet Cores (EPCs) and Evolved UMTS Terrestrial Radio Access Networks (E-UTRAN) as seen in Fig. 11.1. EPCs communicate with each other and with E-UTRANs. EPC contains a Mobile Management Entity (MME) and a System Architecture Evolution Gateway (SGW) together with a Packet Data Network Gateway (PDN GW). E-UTRAN solely contains Evolved Universal Terrestrial Radio Access Network Base Stations (aka eNodeB or eNB) where the User Equipment (UE) communicates with eNB and eNBs communicate with each other and with the EPCs. There is one-to-one communication between UE and eNB but there is one-to-many communication among eNB, MME, and SGW.

Both WiMAX and the UMTS successor technology LTE use Orthogonal Frequency Division Multiplexing (OFDMA) as the core modulation technology on the air interface in downlink direction. TSG-RAN#30 group has found out that initial system level evaluations with 5 MHz allocation, the spectral efficiency improvements achievable with a CDMA-based system according to an “evolutionary” approach and the spectral efficiency improvements with an OFDMA-based system are both attractive. Although using CDMA-based approach might bring smooth migration, OFDMA can avoid the prior constraints and allows for a more free choice of design parameters, making it easier to fulfil the requirements. OFDMA-based downlink physical layer provides more attractive solution when the complexity increases with larger bandwidths and higher MIMO configurations. In uplink direction, however, these two systems differ. While WiMAX uses OFDMA, LTE standardization group has decided to use localized SC-FDMA (see Chap. 7) instead.

In the remainder of this chapter, we first introduce the SAE, starting from EPC and followed by E-UTRAN and UE. Then, we briefly talk about the protocol layers of air interface. Last, we discuss MAC and PHY layers including MIMO support.

11.1 EPS: Evolved Packet System

The EPS (aka SAE) is a "flat", all-IP based core network with a simplified architecture and open interfaces. The EPS is based on TCP/IP protocols to enable PC-like services including voice, video, rich media, and messaging. This migration also enables improved interworking with other fixed and wireless communication networks as seen in Fig. 11.2. The EPS architecture is similar to WiMAX architecture in terms of set of functionalities. The EPS introduces the following entities: MME, SGW, and PDN GW. The control plane functionality in ASN-GW of WiMAX is assigned to MME and data plane functionality is preserved in SGW. The reasoning

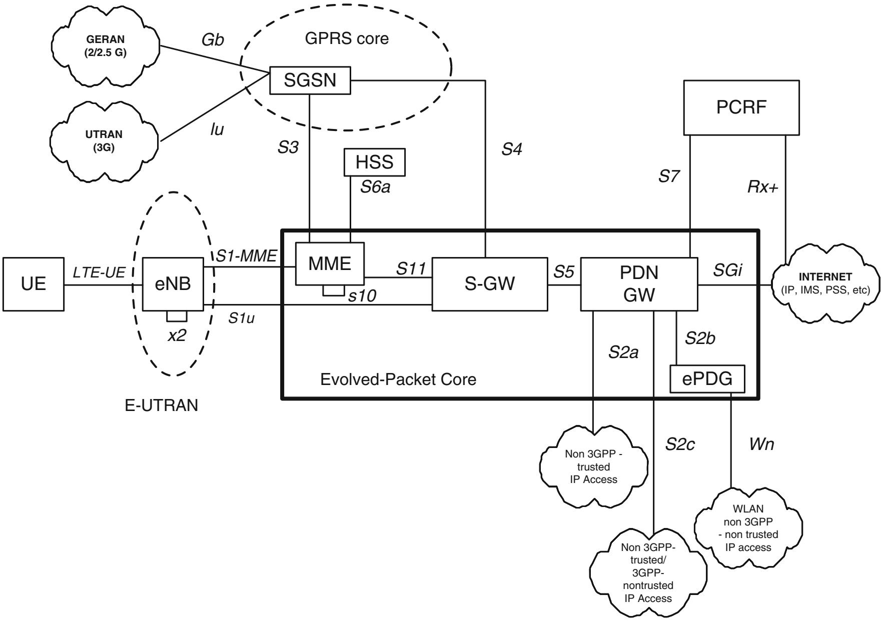


Fig. 11.2 LTE integration

behind that is stated as an optimization of each entity according to its functionality; MME can be optimized for signaling and SGW can be optimized for high bandwidth packet processing. It may give operator to implement those two entities topologically separated or colocated depending on the considered bandwidth latencies and congestion.

11.1.1 MME: Mobility Management Entity

The protocols for mobility management and session management are performed in MME via the Nonaccess Stratum (NAS).[4] The NAS protocols are performed between UE and MME over Stream Control Transmission Protocol (SCTP), which is a connection-oriented transport technology. The NAS consists of EPS Mobility Management (EMM) protocol and EPS Session Management (ESM) protocol: EMM performs the control of mobility and security and EPS is responsible to handle the EPS bearer contexts and control. To perform these functions, there are several interfaces that are defined for MME: MME connects to eNBs with S1-MME (S1-AP) interface and SGWs with S11 interface. The NAS protocols hosted in the MME also specify the procedures for the support of the mobility between LTE and other 3GPP or non-3GPP access networks. For instance, MME communicates with SGSN through S3 interface. Also, part of the NAS protocol is authentication and authorization of UE; consequently, MME retrieves the information from the Home Subscriber Service (HSS) through S6a interface.

In brief, the MME hosts the following functions:

- Selecting SAE GW for a UE at the network entry
- Performing intra-LTE handover
- Paging – distribution of messages to eNBs
- Handing security key management
- Providing mobility in idle state
- Controlling SAE bearer – de/activation
- Ciphering and integrity protection of NAS signaling
- Allocating temporary IDs to UEs
- Handling mobility to other 3GPP or non-3GPP access networks
- Terminating the S6a interface toward the home HSS when UE roams
- Supporting Lawful intercept

EPS introduces an *MME pool* definition in order to facilitate the change of MME during the handover. UEs are supposed to change the MME only when they cross into another pool's service area. The context in the old MME is retrieved through S10 interface by the new MME.

[4] "The protocols between UE and MME that are not terminated in the E-UTRAN."

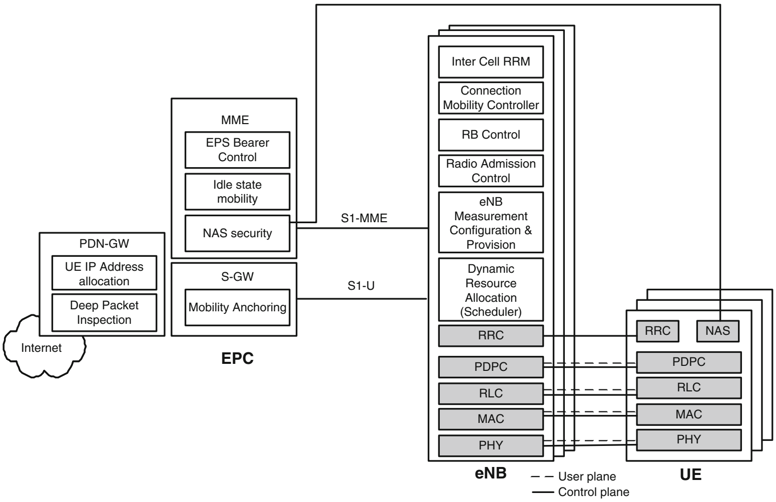


Fig. 11.3 LTE feature distribution

11.1.2 SGW: Serving Gateway

SGW (aka SAE Gateway) is responsible to provide routing and forwarding of user data packets with S1-U interface. SGW connects to PDN GW with S5 interface and gets instruction from MME through S11 interface. SGW is responsible for data paths and handles IP header compression, encryption of user data streams, termination of U-plane packets for paging reasons, and switching of U-plane to support UE mobility as seen in Fig. 11.3. It provides handover function when handover is between LTE and other 3GPP/2 technologies through S4/S103 interface. SGW terminates the data path and triggers paging when UE enters idle mode. SGW is responsible to store the context of UEs. In case of lawful interception, it also performs replication of the user traffic.

Note that there is no communication among SGWs since SGW is designed as an enforcement point. Notifications come from MME to perform the creation/deletion or switching of data plane. If there is data for UE in idle mode, notification goes to MME. The same way if there is an inter-RAN handover then SGW notifies PDN GW to change the data path to a target SGW.

11.1.3 PDN GW: Packet Data Network Gateway

The Packet Data Network Gateway (PDN GW) provides connectivity to external packet data networks and operates as the main mobility point. PDN GW (aka P-GW) is connected to Policy and Charging Rules Function (PCRF) with S7 interface to retrieve policy information. UE may connect to multiple PDN GWs, and UE IP address allocation is performed through PDN GW as well. PDN GW is also first interface toward Internet or hosted services such as IMS, PSS (Packet-Switched Streaming Service), etc, with SGi interface. The PDN GW performs deep packet inspection for a user along with lawful interception. To provide QoS, transport level packet marking for MPLS or DiffServ is performed in PDN GW as well. Charging support for uplink and downlink together with rate enforcement are also provided in PDN GW. The PDN GW is an anchor for mobility between 3GPP and non-3GPP technologies such as WiMAX, 3GPP2 (CDMA 1X and EV-DO), and WLAN through various sets of interfaces seen in Fig. 11.2.

11.2 E-UTRAN

E-UTRAN (Evolved UMTS Terrestrial Radio Access Network) consists of eNBs, which is the base station of LTE and responsible to provide the E-UTRA user plane and control plane. eNBs function as a base station. eNBs are connected to each other with full mesh in order to provide the following functions:

- Transfer of user data in order to provide user data transfer capability across the E-UTRAN between the S1 and LTE-Uu (air) interfaces.
- Radio channel ciphering and deciphering in order to protect transmitted data over the air against unauthorized third party. The key for ciphering and deciphering is derived through signaling or session-dependent information.
- Integrity protection in order to avoid alteration of the transmitted data by an unauthorized third party.
- Header compression in order to provide a compression for a particular network layer or protocol combination such as TCP/IP and RTP/UDP/IP.
- Mobility control functions:
  - Handover manages the radio interface by radio measurements, and it is used to maintain the Quality of Service requested by EPC. It transfers context during handover to target eNB.
  - Paging provides a mechanism to a UE to contact the E-UTRAN when it is in LTE idle state.
  - Positioning is currently being designed to provide physical location information for UE.
- Intercell interference coordination in order to reduce the intercell interference with coordination. This is part of multicell RRM function that considers all information from multiple cells.
- Connection setup and release in order to participate in processing of the end-to-end connection setup and release; maintains and manages the end-to-end connection.
- Load balancing in order to distribute the uneven load distribution to keep the call dropping probabilities minimum. Load balancing may result in handover or cell reselection.

- Distribution function for NAS messages in order to transfer the messages transparently for RRC and S1-AP protocol.
- NAS node selection function in order to select the MME/S-GW for UE.
- Synchronization in order to maintain the timing between different nodes within the network.
- Radio access network sharing in order to provide sharing of radio access network by multiple PLMNs (Public Land Mobile Network). This mechanism directs the UE to appropriate PLMN. E-UTRAN broadcasts the PLMN-ids in the air link (up to 6). The UE selects one of the PLMN-id and notifies E-UTRAN in random access procedure.
- MBMS function in order to ensure the transmission of the same data to multiple recipients.
- Subscriber and equipment trace in order to provide trace of the subscriber equipment. Traces are initiated by core network and a trace setup is transferred on X2 or S1 interface during handover.

11.2.1 eNB: Evolved NodeB

Evolved NodeB (eNB) is the only entity in the evolved-RAN (E-UTRAN) that interfaces with User Equipment (UE) through LTE-UE interface. eNB hosts the PHY, MAC, Radio Link Control (RLC), and Packet Data Control Protocol (PDCP) layers. eNB can support FDD mode, TDD mode, or dual-mode operation with the protocol model depicted in Fig. 11.4.

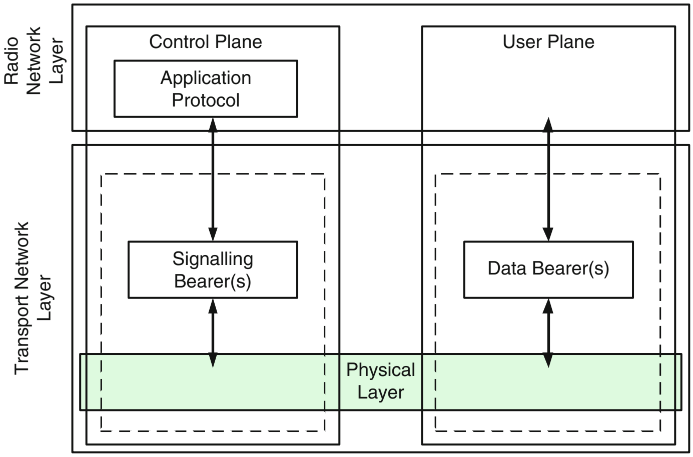


Fig. 11.4 Protocol model of E-UTRAN

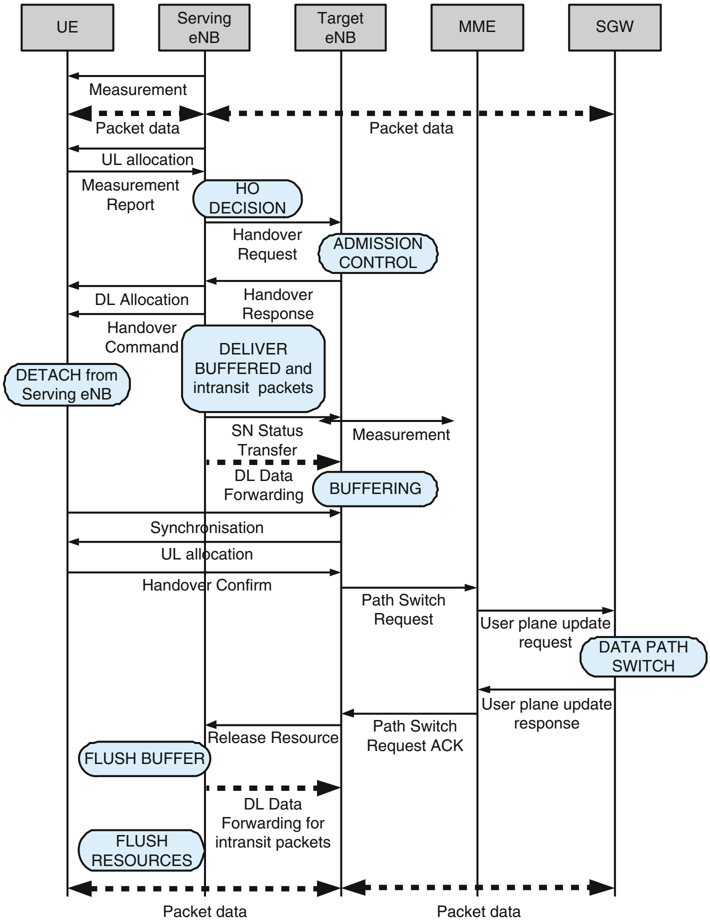


Fig. 11.5 Handover through X2 interface

User plane protocols[5] implement the bearer service in order to carry user data. Control plane protocols control the bearers and the connection between the UE and the network. eNBs are interconnected with each other with X2 interface to support handover of user equipment as seen in Fig. 11.5. The eNBs also communicate to the EPC through S1-flex interface as seen in Fig. 11.1.

S1-flex[6] allows eNB to be connected to multiple set of MMEs and SGWs for redundancy and load sharing. E-UTRAN infrastructure can be shared by more than one operator by having separate MME, SGW, and PDN GW.

[5] The radio interface protocols are defined in documents TS 36.2xx and TS 36.3xx.

[6] The S1 interface protocols are defined in documents TS 36.41x.

11.3 UE: User Equipment

User Equipment (UE) consists of user-plane and control-plane protocol stack. User-plane protocol stack consists of PDCP, RLC, MAC, and PHY layers, which communicates with eNB through LTE wireless link. Control-plane protocol stack contains NAS and RRC in addition to user-plane protocol stack. NAS in UE communicates directly with NAS in MME, and RRC in UE communicates with RRC in eNB as seen in Fig. 11.3.

NAS control-plane protocol is responsible for SAE bearer management, authentication, idle mode mobility/paging handling, and security control.

RRC control-plane protocol is responsible for paging/broadcast, RRC connection management, Radio Bearer (RB) control, mobility functions, and UE measurement and reporting.

E-UTRAN provides two identities for UE: C-RNTI,[7] which is a unique UE identification at cell level to identify RRC connection, and Random value for contention resolution, which is for a transient condition to resolve a contention.

Network level identities include MME identity, which UE presents to the eNB in the idle state, eNB or cell identity, which is given to new eNB in order to have it retrieve UE context, and tracking area id, which describes the paging region. eNB broadcasts cell id, tracking area id, and one or more PLMN ids (identifying each operator).

UE cycles among LTE_DETACHED, LTE_ACTIVE, and LTE_IDLE as seen in Fig. 11.6. During detached mode, UE looks for attachment point to register and it shifts to active mode after registration. In the active mode, it performs its normal operation and initiates handover if needed. During handover, data integrity of packets is satisfied either by buffering in eNB and forwarding to target eNB or bicasting from SAE GW to candidate eNBs. Packet level ordering can be addressed with PDCP sequence numbers. During idle mode, UE is tracked by MME in tracking

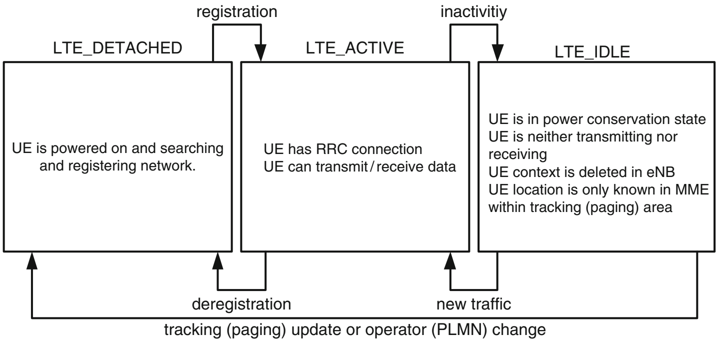

registration
inactivitiy
LTE_DETACHED
LTE_ACTIVE
LTE_IDLE
UE is powered on and searching and registering network.
UE has RRC connection
UE can transmit/receive data
UE is in power conservation state
UE is neither transmitting nor receiving
UE context is deleted in eNB
UE location is only known in MME within tracking (paging) area
deregistration
new traffic
tracking (paging) update or operator (PLMN) change

Fig. 11.6 UE finite state machine

[7] "Radio Network Temporary Identities (RNTI) are used as UE identifiers within E-UTRAN and in signalling messages between UE and E-UTRAN...".

areas, which may contain more than one eNB. And if new traffic arrives to SGW, UE is paged. To enable intertechnology roaming, UE is allowed to respond from different 3GPP technologies such as UMTS/GPRS or UTRAN/GERAN. This list is being tried to be extended for 3GPP2 and IEEE access technologies such as WiMAX and WiFi.

11.3.1 Reference Points

3GPP System Architecture Evolution (SAE) has an objective to migrate the current system to a better technology. Interoperability provides coexistence and colocation with GERAN/UTRAN on adjacent channels. E-UTRAN terminals should support measurements, handover to/from UTRAN or GERAN. The interruption time between E-UTRAN and UTRAN/GERAN should be less than 300ms. Figure 11.2 shows a system architecture possibly relying on different access technologies where WiMAX falls into trusted non-3GPP IP Access.

The interfaces between the SGSN in 2G/3G Core Network and EPC will be based on the GPRS Tunneling Protocol (GTP). The GTP protocol consists of two parts: the GTP-C and the GTP-U. The GTP-C is for control purpose in order to create, modify, and delete the GTP tunnels. Unlike GRE, tunnel creation in GTP is with explicit signaling. The GTP-U transports the user data and some control information. The GTP header is illustrated in Fig. 11.7.

The following interfaces are defined:

LTE-Uu: Reference point of the radio interface between UE and eNB.

S1-MME: Reference point for control plane between E-UTRAN and MME. S1-MME uses SCTP as the transport protocol.

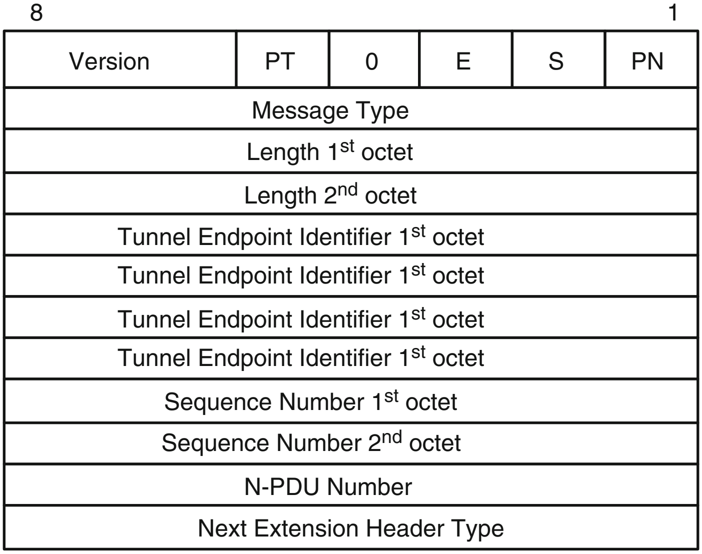
S1-U: Reference point between E-UTRAN and SGW for the per-bearer user plane tunneling and inter-eNB path switching during handover with GTP-U as transport protocol.

S2a: Reference point for user plane with related control and mobility support between trusted non-3GPP IP access and the gateway based on Proxy Mobile IP. S2a also supports Client Mobile IPv4 FA mode if PMIP is not available.

S2b: Reference point to provide the user plane with related control and mobility support between evolved Packet Data Gateway (ePDG) and the PDN GW over PMIP.

S2c: Reference point to provide the user plane with related control and mobility support between UE and the PDN GW. This reference point is implemented over trusted and/or untrusted non-3GPP access and/or 3GPP access over CMIP colocated mode.

S3: Reference point between SGSN and MME to enable user and bearer information exchange for inter-3GPP access network mobility in idle and/or active state over GTP and Gn reference point as defined between SGSNs.



8					1
Version	PT	0	E	S	PN
Message Type					
Length 1st octet					
Length 2nd octet					
Tunnel Endpoint Identifier 1st octet					
Tunnel Endpoint Identifier 1st octet					
Tunnel Endpoint Identifier 1st octet					
Tunnel Endpoint Identifier 1st octet					
Sequence Number 1st octet					
Sequence Number 2nd octet					
N-PDU Number					
Next Extension Header Type					

Fig. 11.7 GTP header: the version number determines the version of this header and GTP is backward compatible. The PT bit stands for protocol type to identify whether this is standard GTP or GTP′, which is used for charging purposes. The E bit stands for extension header and the S bit is for sequence number. The N-PDU number bit indicates whether there is an N-PDU number. Message type indicates the type of the message such as echo request, node alive request, create/deleted PDP context request, sending routing information, etc. The length field indicates the length of the payload and the TEID identifies the tunnel end points

S4: Reference point to provide the user plane with related control and mobility support between SGSN and the SGW over GTP and Gn reference point as defined between SGSN and GGSN.

S5: Reference point to provide user plane tunneling and tunnel management between SGW and PDN GW when SGW relocation is needed due to UE mobility and if the SGW needs to connect to a noncollocated PDN GW for the required PDN connectivity. GTP and the IETF-based PMIP are possible solutions.

S6a: Reference point to enable transfer of subscription and authentication data for authenticating/ authorizing user access to the evolved system (AAA interface) between MME and HSS.

S6c: Reference point between PDN GW, Home PLMN[8] (HPLMN), and 3GPP AAA server for mobility-related authentication if needed.

S6d: Reference point between SGW, Visited PLMN[9] (VPLMN), and 3GPP AAA Proxy for mobility-related authentication if needed.

[8] Home Public Land Mobile Network.

[9] Visited Public Land Mobile Network (GSM).

S7: Reference point to provide transfer of QoS policy and charging rules from PCRF to Policy and Charging Enforcement Function (PCEF) in the PDN GW over Gx interface.

S8a: Reference point based on GTP protocol and the Gp interface defined between SGSN and GGSN. It is for home-routed traffic in order to provide user plane with related control between the SGW in VPLMN and the PDN GW in HPLMN. S8a is a variant of S5 for roaming and S8b is available that supports PMIP.

S9: Reference point between hPCRF and vPCRF used in roaming to enforce in the VPLMN of dynamic control policies from the HPLMN.

S10: Reference point between MMEs for MME relocation and MME to MME information transfer.

S11: Reference point between MME and SGW to instruct the decisions for enforcement point.

SGi: Reference point between the PDN GW and the packet data network. Packet data network may be an operator-external public or private packet data network or an intraoperator packet data network. This reference point corresponds to Gi for 2G/3G accesses.

Rx+: Reference point between the Application Function and the PCRF defined in the 3GPP TS 23.203.

Wn*: Reference point between the untrusted non-3GPP IP access and the ePDG. Traffic on this interface has to be forwarded toward ePDG.

11.4 System Aspects

11.4.1 QoS

EPS bearers are defined as an aggregate point of one or more IP flows. The GTP bearers exist between UE and the PDN GW as seen in Fig. 11.8. Aggregation is

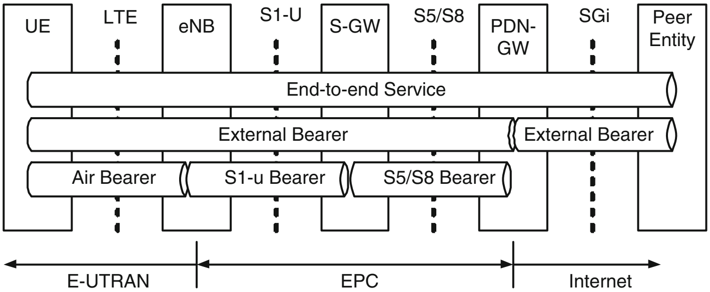


Fig. 11.8 Bearer: GTP tunnel IDs over S5/S8a interfaces have a one-to-one mapping to S1 interface tunnel IDs as well as Radio Bearer IDs over the Radio Bearer

defined as the convenance to perform the same type of packet forwarding for same type of flows, otherwise one flow would require one EPS bearer. Binding is performed in UE for uplink and in PDN GW for downlink packets. The SAE Radio Bearer Service is responsible for the transport of the SAE Bearer Service data units between eNB and UE according to QoS profile. The SAE Access Bearer Service is responsible for the transport of the data units between SGW and eNB according to QoS profile. The SAE Access Bearer links to SAE Radio Bearer, which links to a logical channel.

There are default bearers and dedicated bearers. The default bearer is started at the time of start-up to carry all traffic. The dedicated bearers, on the other hand, carry IP traffic with differentiated forwarding. The default bearer is a nonguaranteed bit rate (non-GBR) bearer, which can suffer from packet losses. A dedicated bearer with GBR has a guaranteed bit rate and Maximum Bit Rate (MBR) parameters. Also, GBR bearers that belong to the same UE share an Aggregate Maximum Bit Rate (AMBR).

QoS is based on label and Allocation and Retention Priority (ARP) where label maps the defined characteristics to specific labels and ARP decides whether or not the bearer can be accepted based on resource availability.

11.4.2 Security

SAE/LTE security defines a new architecture with extended key hierarchy. It prohibits SIM (Subscriber Identity Module) access but uses USIM (Universal Subscriber Identity Module) from Rel-99. This is the master key (128 bits) and there is possibility to add 256-bit keys later.

The subscriber authentication is through AKA procedure between the UE and the MME. Additional Access Security Management Entity (ASME), colocated with MME, is introduced to protect the NAS signaling (encryption and integrity via AES and SNOW 3G as the selected crypto-algorithms).

The keys used for NAS protection are separate in eNB and EPC. This makes it impossible to use the eNB key in order to extract the EPC key. CK/IK keys are confined to home network and ASME receives the derived key (K_{ASME}) for authentication with the UE as seen in Fig. 11.9. ASME passes this key to MME and also sends keys to eNB derived from K_{ASME}. MME retains the keys when UE goes to idle state.

When UE enters the connected state, eNB keys are sent to eNB from EPC. If keys are detected to be corrupted, UE restarts the radio attachment procedure. *NAS* and K_{eNB} keys are derived from K_{ASME}, which never leaves the EPC. From K_{eNB}, eNB and UE derive the UP and *RRC* keys. These three keys are deleted when UE goes to idle or null state. NAS keys are used for the protection of NAS traffic, UP is used for the protection of U-Plane traffic, and *RRC* key is used only for protection of *RRC* traffic.

Extended key hierarchy allows fast key refreshing for intra-LTE handovers. Key context is shared among eNBs during handover. For handovers between

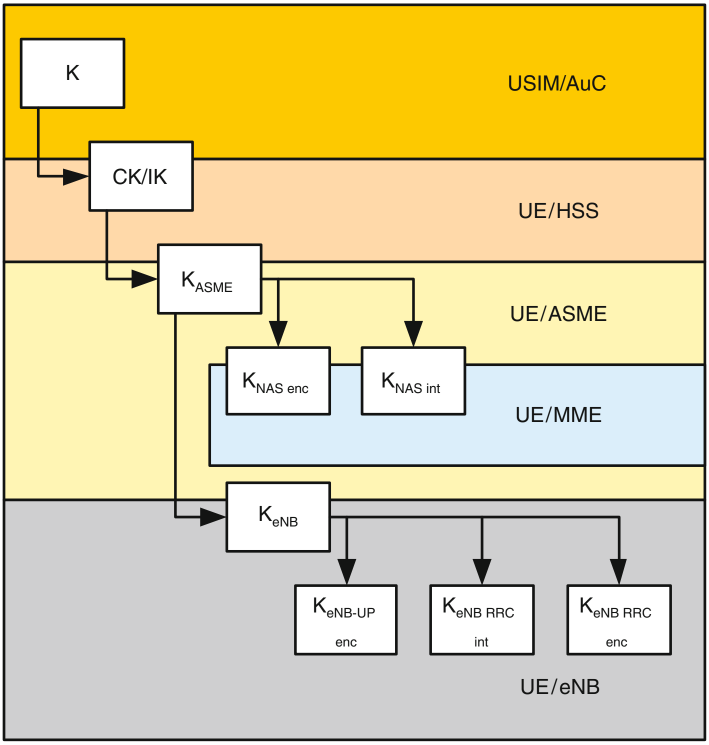


Fig. 11.9 Key hierarchy; USIM: Universal Subscriber Identity Module, AuC: Authentication Center

E-UTRAN and 2G/3G systems, key exchange is between SGSN and MME. For UTRAN/GERAN to E-UTRAN handovers, SGSN sends CK/IK to MME to derive K_{ASME} but in the opposite direction, MME derives CK/IK from K_{ASME} and sends it to SGSN.

11.5 LTE Higher Protocol Layers

Protocol layer architecture for LTE system is shown Fig. 11.3. There are control plane and user plane protocol stacks in eNB and UE.

Protocol layers in UE are as follows:

- NAS: Nonaccess Stratum
- RRC: Radio Resource Control
- PDCP: Packet Data Convergence Protocol
- RLC: Radio Link Control
- MAC: Medium Access Control
- PHY: Physical Layer

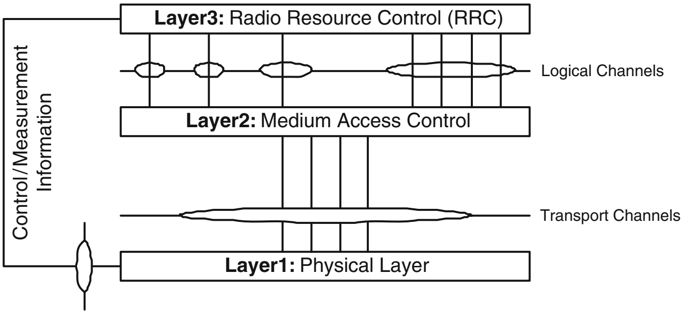

Layer3: Radio Resource Control (RRC)
Logical Channels
Layer2: Medium Access Control
Transport Channels
Layer1: Physical Layer
Control/Measurement Information

Fig. 11.10 Channel structure

Protocol layers in eNB are as follows:

- RRC: Radio Resource Control
- PDCP: Packet Data Control Protocol
- RLC: Radio Link Control
- MAC: Medium Access Layer
- PHY: Physical Layer

Communication between these layers is established via channels as seen in Fig. 11.10. First, let us look at the channel structure and then delve into protocol layers.

11.5.1 Communication Channel Structure

Each channel is characterized with a certain set of parameters and functions. There are logical channels that are mapped to transport channels. And, there are transport channels that are mapped to physical channels. Logical channels are identified with respect to the information carried by them and transport channels are distinguished according to their transmission characteristics. Similarly, physical channels are characterized by their configuration for data protection. Fig. 11.11 shows the mapping structure of channels for uplink and downlink.

There are two types of logical channels: control and traffic channels. Control channels are as follows:

- *BCCH:* Broadcast Control Channel is to transmit broadcasting system control information.
- *PCCH:* Paging Control Channel is to transmit paging information when UE is unlocated.
- *CCCH:* Common Control Channel is used by UE when UE has no RRC connection.

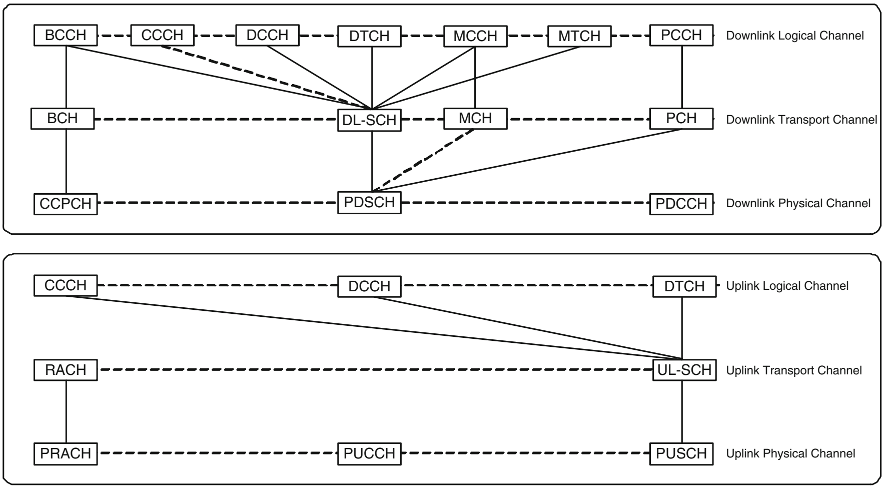


Fig. 11.11 Downlink and uplink channel mapping: dotted lines are still being studied by 3GPP

- *MCCH:* Multicast Control Channel is used to transmit MBMS control information, which is point-to-multipoint (eNB to UEs) and only UEs that receive MBMS use it.
- *DCCH:* Dedicated Control Channel is a point-to-point bidirectional channel used by UE for RRC connection.

Traffic channels are as follows:

- *DTCH:* Dedicated Traffic Channel is a point-to-point bidirectional channel dedicated to one UE to transfer user information.
- *MTCH:* Multicast Traffic Channel is a point-to-multipoint channel for transmitting traffic data from the network to the UE.

Transport channels provide structure passing data to/from higher layers, mechanism to configure PHY, status indicators to higher layers (CQI, error, etc.) and higher-layer peer-to-peer signaling. Transport channels for downlink are as follows:

- *BCH:* Broadcast Channel transmits entire cell area with fixed transport format.
- *DL-SCH:* Downlink Shared Channel is used for HARQ, dynamic link adaptation, UE DRX (discontinuous receive) for power save, dynamic and semistatic allocation, and beamforming.
- *PCH:* Paging Channel is used for UE DRX, broadcast over cell coverage.
- *MCH:* Multicast Channel provides support for Multicast Broadcast -Single Frequency Network (MB-SFN) with semi-static resource allocation.

Transport channels for uplink are as follows:

- *UL-SCH:* Uplink Shared Channel is for HARQ, dynamic link adaptation, support for UE DRX, and dynamic and semistatic resource allocation with 1/3 turbo coding.

- *RACH:* Random Access Channel is for limited control information, which has collision risk.

Transport channels are connected to physical channels. LTE downlink physical channels are as follows:

- *PDSCH:* Physical Downlink Shared Channel is utilized for data and multimedia transport. It is designed for high data rates with QPSK, 16QAM, and 64QAM modulation with 1/3 turbo coding and spatial multiplexing.
- *PDCCH:* Physical Downlink Control Channel conveys UE-specific information with QPSK-only modulation for robustness. Up to first three OFDM symbols in the first slot of a subframe are used for PDCCH.
- *CCPCH:* Common Control Physical Channel conveys cell information with QPSK-only modulation and convolutional coding. CCPCH is sent close to center frequency.

LTE uplink physical channels are the following:

- PUSCH: Physical Uplink Shared Channel is allocated subframe basis by the UL scheduler with QPSK, 16QAM, or 64QAM modulation.
- PUCCH: Physical Uplink Control Channel carries uplink control information including CQI, ACK/NACK, HARQ, and uplink scheduling requests.

Each physical channel has a defined algorithm for bit scrambling, modulation, layer mapping, precoding, and scheduling. Also, note that layer mapping and precoding are applicable when MIMO mode is present.

11.5.2 NAS Layer

NAS layer is used between UE and MME for establishing control signaling. This communication is used for the following:

- Network entry (attach)
- Authentication
- Data bearers setup
- Mobility management

The NAS signalling security is provided by ciphering and integrity protection. Transfer of NAS messages from/to UE is handled by RRC layer.

11.5.3 RRC Layer

The Number of RRC states in LTE is reduced to 2 as compared to that in predecessors. Two states of RRC are RRC_IDLE and RRC_CONNECTED. RRC may perform one of the listed functions in Fig. 11.12 according to these states.

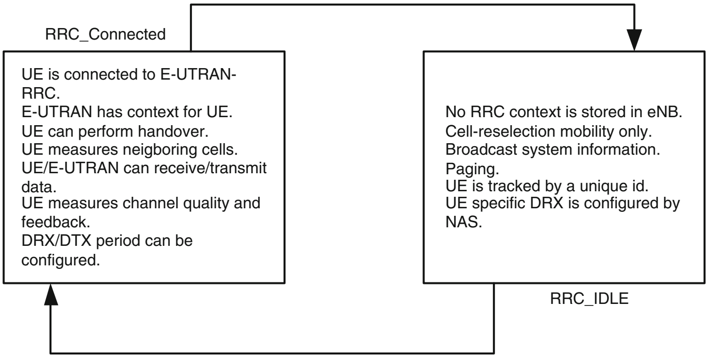

RRC_Connected
UE is connected to E-UTRAN-RRC.
E-UTRAN has context for UE.
UE can perform handover.
UE measures neigboring cells.
UE/E-UTRAN can receive/transmit data.
UE measures channel quality and feedback.
DRX/DTX period can be configured.
No RRC context is stored in eNB.
Cell-reselection mobility only.
Broadcast system information.
Paging.
UE is tracked by a unique id.
UE specific DRX is configured by NAS.
RRC_IDLE

Fig. 11.12 UE states

The RRC layer of eNB is responsible to broadcast system information, perform paging, and establish an RRC connection with UEs to allocate temporary identifiers (RA-RNTI). The RRC layer also configures the signaling radio bearer for RRC connection and is responsible for integrity of RRC messages.

RRC plays a key role in mobility. UE measurement and reporting, intra-LTE handover, UE cell re/selection, and context transfer are handled by the RRC layer. RRC also facilitates MBMS services.

11.5.4 PDCP Layer

Figure 11.13 illustrates the PDCP, RLC, and MAC layers in detail for both eNB and UE. The PDCP layer provides ROHC header compression and decompression for efficient air bandwidth usage. It transfers the PDCP SDU received from NAS to RLC layer and vice versa. It supports ciphering of user plane and control plane data. PDCP PDU comprises PDCP header and PDCP SDU.

11.5.5 RLC Layer

The RLC layer is responsible to transfer traffic PDUs between UE and eNB with segmentation if needed and applies error correction through ARQ for received data. It applies concatenation, in-sequence delivery, and duplicate detection. RLC PDU comprises RLC header and RLC SDU. RLC layer provides three different reliability modes:

AM: Acknowledge Mode requires acknowledgement and is good for unreal time services such as file download.

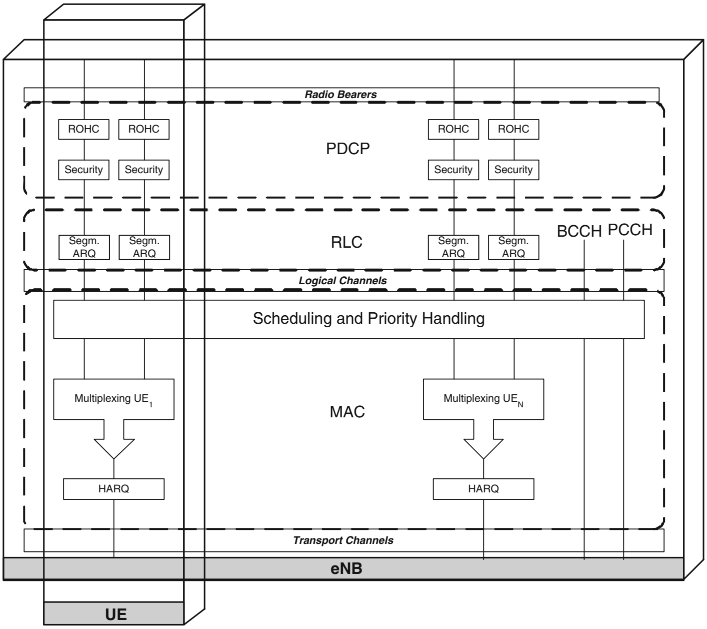


Fig. 11.13 Layer and channel structure for UE and eNB

UM: Unacknowledge Mode does not require an acknowledgement and is suitable for real time services such as video streaming.

TM: Transparent Mode implement implicits acknowledgement and is used when file sizes are known as in broadcasting.

11.6 LTE MAC layer

The MAC layer services and functions include mapping between logical channels and transport channels. It multiplexes the RLC PDUs into transport blocks or demultiplexes the RLC PDUs from transport blocks in the reception side. Measurement reporting for traffic and error correction through HARQ is also done in MAC layer. MAC layer's main function is scheduling that differentiates between logical channels and different UEs.

11.6.1 Scheduling

The eNB has a scheduler to control the time/frequency resources for a given time for uplink and downlink. The scheduler dynamically allocates resources to UEs at each TTI (Transmission Time Interval) via the C-RNTI on L1/L2 channel(s). A UE always monitors the L1/L2 control channel(s) to find possible allocation. Predefined resource allocation is also possible where UE is notified by the configuration and then the allocation is done without C-RNTI where UE blindly does the decoding.

Depending on the channel conditions, scheduler selects the best multiplexing for UE. The decision can be based on any combination of the following:

- QoS parameters
- Measurements
- Buffered payloads
- Pending retransmissions
- CQI reports from the UEs
- UE capabilities
- UE sleep cycles
- Measurement gaps/periods
- System parameters such as bandwidth and interference level/patterns

Downlink LTE considers the following schemes as a scheduler algorithm:

FSS: *Frequency Selective Scheduling* assigns transmission resources to a user using the selective resource blocks to offer the best performance. Channel-side information is necessary for FSS and it can increase capacity over TDS (Time Domain Scheduling). This type of scheduling may be used at higher speeds, at cell edges, or for low overhead services.

FDS: *Frequency Diverse Scheduling* does not require channel-side information since it assigns resources distributed across the transmission bandwidth.

PFS: *Proportional Fair Scheduling* is the preferred scheduling mechanism, which basically is a C/I scheduler but with a delay component to address the delay-sensitive traffic.

Link adaptation is performed through adaptive modulation and coding. A user uses the same coding and modulation for all PDUs; however, in case of MIMO, different streams may use different modulation and coding.

11.6.2 HARQ

HARQ framework in LTE considers incremental redundancy and special case of chase combining.

HARQ can be synchronous or asynchronous. Synchronous HARQ requires that transmission occurs at known time instants. No explicit signaling is required; on the other hand, for asynchronous HARQ, explicit signaling is required to accommodate HARQ process that happens anytime. HARQ can also be adaptive or nonadaptive; adaptive HARQ has the ability to change the modulation, resource block allocation, and duration of transmission.

Note that synchronous operation requires less control signaling and has significant advantage when it is nonadaptive since soft-combining can be performed. This mode is selected for uplink. However, in the downlink, asynchronous and adaptive HARQ mode is considered.

11.6.3 Cell Search

UE performs cell search in order to acquire cell id and time and frequency synchronization. Synchronization Channel (SCH) and Broadcast Channel (BCH) are detected during the cell search; SCH is for timing information such as symbol timing and frequency of the downlink signal; BCH is to broadcast certain set of cell-specific information such as transmission bandwidth, cell id, antenna configuration, etc.

Within a frame time at least one SCH and BCH are transmitted. In frequency domain, central portion is allocated for SCH and BCH transmissions. This is because UE first detects the central part of the spectrum regardless of the receiving bandwidth capability of the UE and the transmission bandwidth of eNB.

There are primary and secondary SCHs. P-SCH and S-SCH are transmitted on subframe 0 and 5 by two symbols. With use of P-SCH, UE detects the carrier frequency, SCH symbol timing, and cell id. Then, with the use of S-SCH, UE detects the radio frame timing, cell id group, MIMO antenna configuration used for BCH, and CP length. BCH is also divided into primary and dynamic portions: primary portion gives information to decode the dynamic portion, which contains system information fields.

11.6.4 Power Control

Power control is considered to mitigate path loss and shadowing. Power control for uplink consists of open and closed-loop schemes to control energy per resource element applied for a uplink transmission. Within a cell, closed-loop power control adjusts a set point determined by the open-loop power control component. eNB can instruct UE with a periodic transmit power command in order to have UE to adjust its transmit ERPE (Energy per Resource Element) where resource element energy is defined as the energy prior to CP insertion.

11.6.5 Intercell Interference Mitigation

Intercell interference mitigation is proposed to be handled via three different approaches: randomization, cancelation, and coordination and avoidance.

There are couple of techniques to cite for randomization: cell-specific scrambling, which is random scrambling after channel coding/interleaving and cell-specific interleaving, which is also known as Interleaved Division Multiple Access (IDMA). Also, cancelation can be performed with spatial suppression or detection of intercell interference. UE needs to be signaled to find out whether it can cancel the ICI; eNB informs the interfering signal configuration (interleaver pattern ID, modulation scheme, FEC scheme, and coding rate) to the UE through synchronization with other eNBs.

Coordination and avoidance is a restriction-based intercell interference remedy technique. Downlink resource management is coordinated between cells. These restrictions are based on available time/frequency resources. Depending on the configuration (static or semistatic) intercell communication is needed to reconfigure the resources on a time scale. Additional UE measurements may be needed as well in every 100 ms.

11.6.6 Internode B Synchronization

Inter-eNB synchronization can be done by GPS or through another eNB. GPS provides an absolute time reference, which has a signal period of 2.56s. This means that for every 6,400 frames the start of a 256-frame period coincides with an integer GPS second.

11.6.7 Physical Layer Measurements

The UE reports the channel quality to the eNB. Time granularity of CQI is important to efficiently trade-off between overhead and link adaptation/scheduling performance. Measurements are for two purposes: scheduling and mobility. For scheduling, measurements are used to apply selective scheduling, selection of modulation/coding scheme, interference management, and power control. Measurements for mobility are used to detect the suitable cells and decide for handover.

11.6.8 Evolved-Multicast Broadcast Multimedia Services

LTE specification will include support for Evolved Multimedia Broadcast Multicast Services (E-MBMS), which is a multimedia service performed either with a single cell broadcast or multicell mode [aka MBMS Single Frequency Network (MBSFN)]. In a MBSFN synchronization area, all eNBs are synchronized to perform MBSFN transmission as seen in Fig. 11.14. In MBSFN area, simulcast

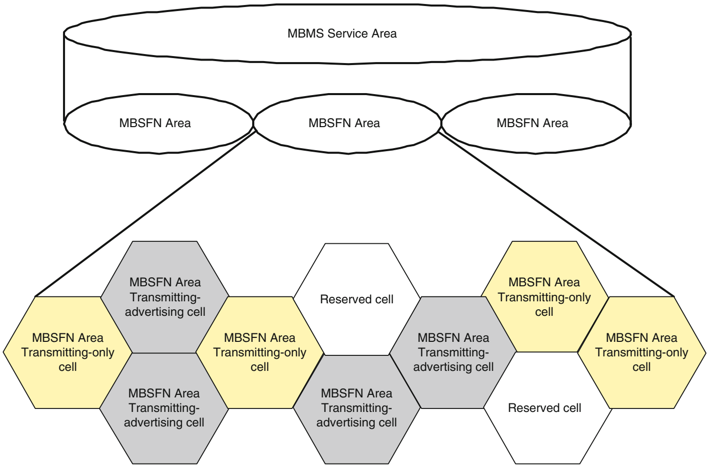

MBMS Service Area
MBSFN Area
MBSFN Area
MBSFN Area
MBSFN Area Transmitting-advertising cell
Reserved cell
MBSFN Area Transmitting-only cell
MBSFN Area Transmitting-only cell
MBSFN Area Transmitting-only cell
MBSFN Area Transmitting-advertising cell
MBSFN Area Transmitting-only cell
MBSFN Area Transmitting-advertising cell
MBSFN Area Transmitting-advertising cell
Reserved cell

Fig. 11.14 E-MBMS network

transmission from multiple cell is performed. Each cell transmits the identical waveform. In the MBSFN area, there are three types of cells: transmitting and advertising cells, transmitting-only cells, and reserved cells. Transmitting and advertising cells are allowed to transmit as well as advertise the MBS services. Transmitting-only cells are restricted to transmission only, and reserved cells does not participate in MBS but other forms of transmission.

MBMS traffic uses DL-SCH in the single cell mode and in the multicell mode each participating cell is time-synchronized to simultaneous broadcast of identical transmission on a common frequency. UE can do over-the-air combining as in macrodiversity with no additional receiver complexity. As a result, UE receives signals with improved SINR.

Delay of arrival between two cells could be quite large if a user is close to a base station. For this reason, subcarrier spacing is reduced to 7.5 KHz and longer CP is used so that UE can combine transmission from different eNBs.

The overall architecture includes E-MBMS gateway and a server as seen in Fig. 11.15. A SYNC protocol is designed in both E-MBMS gateway and the eNBs for synching the same content for a specific time/frequency. E-MBMS is responsible to distribute the content to eNBs with unicast or multicast.

11.6.9 Self Configuration

Dynamic configuration of S1-MME interface is being designed in order to ramp up eNBs efficiently and scalably. Certain prerequisites need to be fulfilled such as providing a remote IP end point to be used for SCTP initialization. Once SCTP

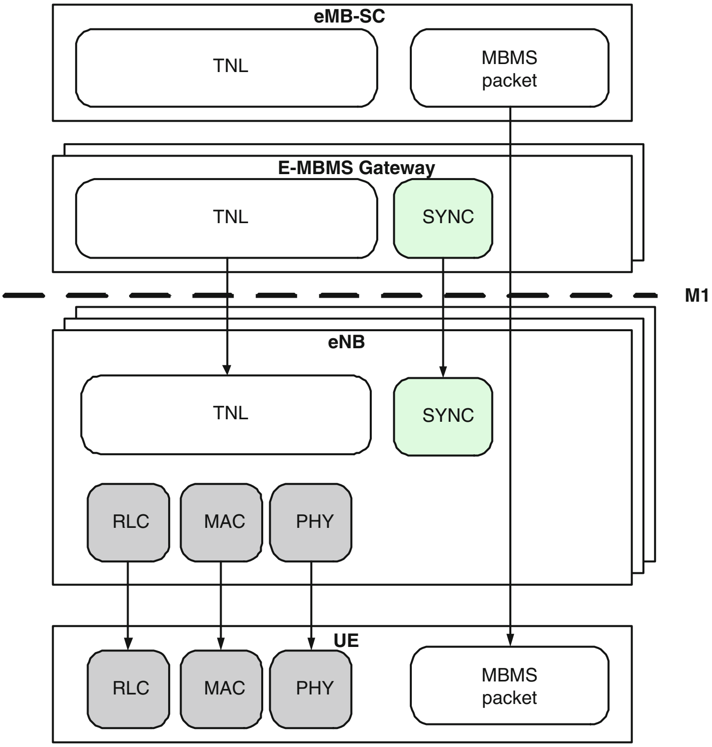


Fig. 11.15 E-MBMS architecture

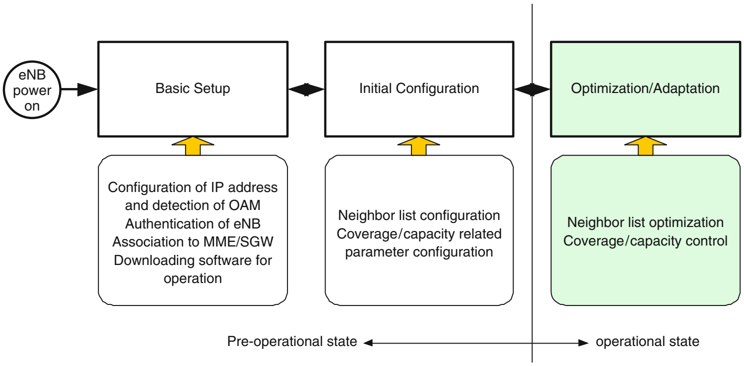

eNB power on
Basic Setup
Initial Configuration
Optimization/Adaptation
Configuration of IP address and detection of OAM
Authentication of eNB
Association to MME/SGW
Downloading software for operation
Neighbor list configuration
Coverage/capacity related parameter configuration
Neighbor list optimization
Coverage/capacity control
Pre-operational state
operational state

Fig. 11.16 Self configuration

is established, through S1-MME, configuration data are exchanged. When the dynamic configuration is completed, S1-MME interface is operational. Figure 11.16 illustrates the procedures for self configuration, which also includes self optimization in which eNBs are optimized adaptively with network statistics.

11.7 LTE PHY Layer

The PHY layer offers data transport to higher layers. The PHY layer[10] is being designed to perform the following functions:

- Error detection on the transport channel and indication to higher layers
- FEC encoding/decoding of the transport channel
- Hybrid ARQ soft-combining
- Rate matching
- Mapping of the coded symbols to physical channels
- Power weighting of physical channels
- Modulation and demodulation
- Frequency and time synchronization
- Radio characteristics measurements and indication to higher layers
- MIMO/transmit diversity beamforming support
- RF processing

The PHY layer is based on OFDMA with a cylic prefix in downlink and on SC-FDMA with a cylic prefix in the uplink. Three duplexing modes are supported: full duplex FDD, half duplex FDD, and TDD. Typical channel bandwidths associated with respective duplexing modes are listed in Table 11.2. Note that transmission bandwidth is defined in terms of resource blocks as seen in Fig. 11.17.

11.7.1 LTE Frame

Frame structure Type-1 shared by both full and half-duplex FDD is presented in Fig. 11.18.[11] Some highlights of LTE frame are as follows:

Table 11.2 Transmission bandwidth configuration in terms of number of Resource Blocks for E-UTRA channel bandwidths

BWMHz	1.4	1.6	3	3.2	5	10	15	20
FDD Type-1	6	N/A	15	N/A	25	50	75	100
FDD Type-2	[6]	TBD	[15]	TBD	25	50	75	100
TDD Type-2	[TBD]	[7]	[TBD]	[16]	25	50	75	100

[10] Structure of the documentation for PHY layer is as follows:

36.300 Overall description
36.211 Physical channels and modulation
36.212 Multiplexing and channel coding
36.213 Physical layer procedure
36.214 Physical layer measurements,

where series can be found in
`http://www.3gpp.org/ftp/Specs/html-info/36-series.htm`

[11] "Frame structure Type-2 is applicable to TDD, which consists of two half-frames of length $T_f = 153600 \times T_s = 5ms$ each as seen in Fig. 11.19. Each half-frame consists of eight slots of

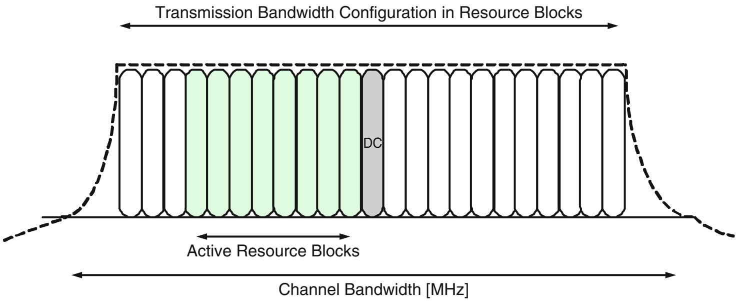


Fig. 11.17 Channel bandwidth and transmission bandwidth representation

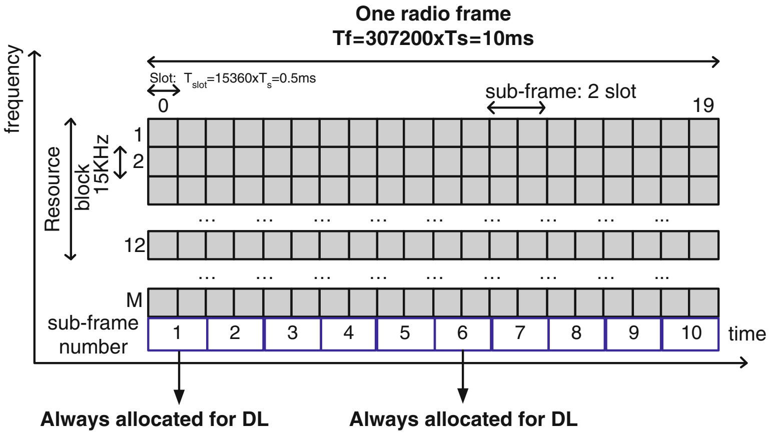


Fig. 11.18 Frame structure type-1 OFDMA/SC-FDMA

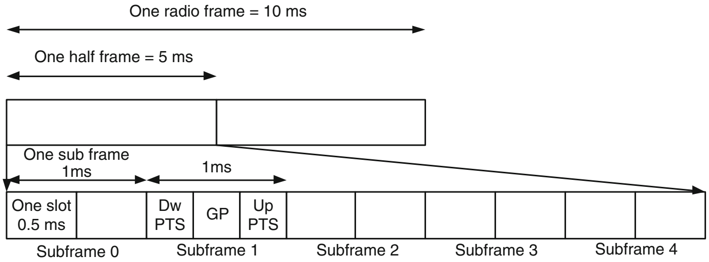


Fig. 11.19 Frame structure type-2 OFDMA/SC-FDMA

length $T_{slot} = 15360T_s = 0.5\,ms$ and three special fields are reserved for DwPTS, Guard Period, and UpPTS. Total length of three fields is 1 ms but each size is configurable. Subframes 1 and 6 consist of DwPTS, GP, and UpPTS and all other subframes are defined as two slots. Subframes 0 and 5 and DwPTS are always reserved for downlink. Switch points can be at 5 or 10 ms. For 5-ms switch point, UpPTS and subframes 2 and 7 are reserved for uplink."

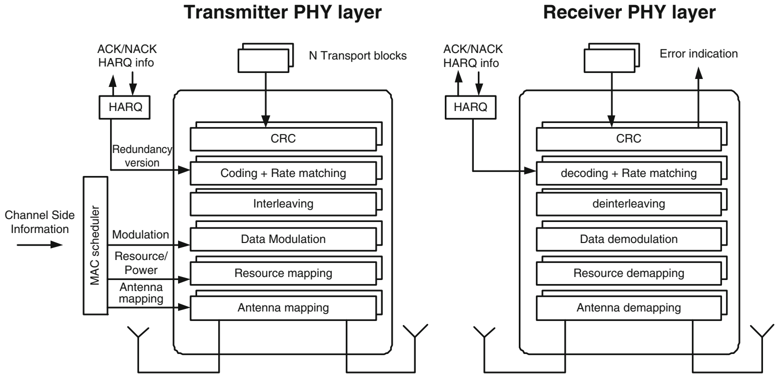


Fig. 11.20 PHY layer interaction: notice that transmitter is eNB and receiver is UE for downlink. For uplink, transmitter is UE and receiver is eNB

- The radio frame structure type-1 has 20 slots with a duration of 0.5 ms where frame length is 10 ms.
- Two adjacent slots constitute a subframe of length 1ms.
- Resource block spans 12 or 24 subcarriers over a slot with a subcarrier bandwidth of 15 KHz or 7.5 KHz, respectively.
- Modulation schemes supported are QPSK, 16QAM, 64QAM.
- Broadcast channel only uses QPSK.
- Turbo coding rate of R = 1/3 and Trellis termination are used.
- CRC-24 is used for error detection.
- Maximum information block size is 6,144 bits.
- Scrambling and channel interleaving are supported.
- Layer mapping and precoding are used for MIMO.

11.7.2 Channel Coding

Figure 11.20 describes the cross layer physical layer model together with MAC interaction. Higher layer data are passed to/from the physical layer in each TTL (time-to-time). After CRC check, channel coding is performed with regard to transport block size, modulation scheme, and resource assignment. Hybrid ARQ has a control over coding and rate matching depending on the HARQ type. Also, MAC scheduler involves in data modulation step since mapping data to resource control blocks is directly controlled by MAC scheduler through L-2 controlled resource assignment. MAC scheduler also involves in physical layer processing and multi-antenna processing since it configures mapping of assigned resource blocks to the available antenna ports.

Transmitter transports signaling of transport format and resource allocation to the receiver. Transmitter can multiplex this information with the HARQ information, and this information, is only used in PHY layer of receiver. In the receiver side, the operation is reversed.

- *Code block segmentation and code block CRC attachment:* There are two polynomials for CRC-24 and one polynomial for CRC-16. The bits before and after CRC attachment are denoted by a_i and b_i, respectively where $B = A + L$ and L is length of the parity bits. $b_i = a_i$ is true for bits from 0 up to $A - 1$ and $b_i = p_{i-A}$ is true for bits between A and $A + L$. If B is larger than the maximum code block size $Z = 6{,}144$ *bits*, block is segmented and additional CRC-24 is performed for each code block. If $B < 40$ then filler bits are added. The output bits are denoted by c_i^r now where r is code block number.
- *Channel coding of data and control information:* The c_i^r are fed to encoder and resultant encoding bits are denoted by d_i^j where j indexes the encoder output stream. Tail biting convolutional coding and turbo coding are applicable to transport channels. Turbo coding with rate 1/3 is used in UL-SCH, DL-SCH, PCH, MCH and tail biting convolutional coding with rate 1/3 with constraint length 7 is used for BCH.

  – *Tail biting Convolutional coding:* Tail biting method uses the last six information bits in the input stream as the initial value of the shift register in order to have the initial and final stages of the shift register the same. The encoder output streams d_j^1, d_j^2, and d_j^3 for given c_k are seen in Fig. 11.21.

  – *Turbo coding:* Parallel Concatenated Convolutional Code (PCCC) is the underlying scheme for turbo coding along with two 8-state constituent encoders and one turbo code internal interleaver as seen in Fig. 11.22. The transfer function of the 8-state constituent code for the PCCC is $G(D) = [1, g_1(D)/g_0(D)]$

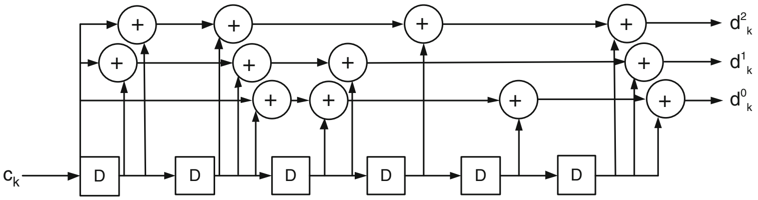


Fig. 11.21 Tail biting convolutional encoder structure of LTE with rate 1/3

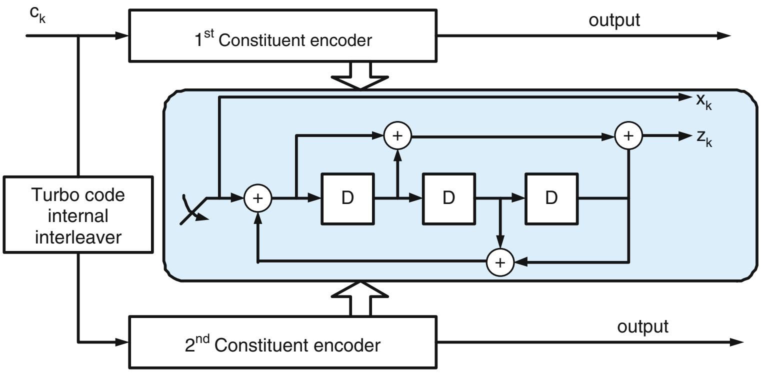

c_k
1^{st} Constituent encoder
output
x_k
z_k
+
D
Turbo code internal interleaver
2^{nd} Constituent encoder
output

Fig. 11.22 Turbo coding structure of LTE with rate 1/3

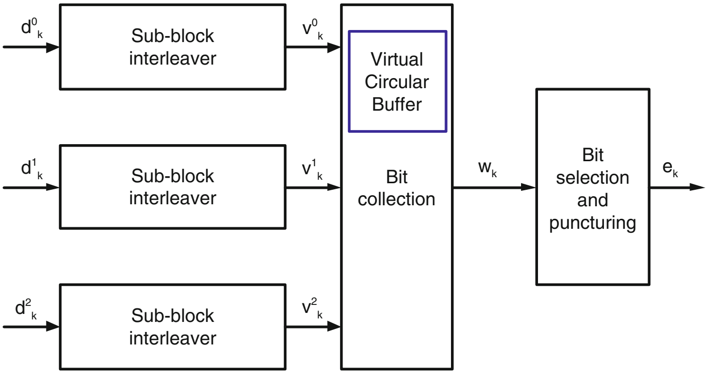


Fig. 11.23 Rate matching structure of LTE

for given $g_0(D) = 1 + D^2 + D^3$ and $g_1(D) = 1 + D + D^3$. The coding rate of turbo encoder is 1/3 as seen in Fig. 11.22. The bits output from the first and second 8-state constituent encoders are denoted by z_j and z'_j, respectively. The bits output from the turbo code internal interleaver are denoted by c'_j. Trellis termination is performed by taking the tail bits from the shift register feedback after all information bits are encoded.

- *Rate matching:* After encoding, rate matching can be applied as seen in Fig. 11.23. The bit streams d^0_k, d^1_k, and d^2_k are interleaved and resultant sequences are v^0_j,v^1_j,v^2_j where $j \in \{0, 1, \cdots, K\}$. Subblock interleaver is a block interleaver, which is followed by the collection of bits and the generation of a circular buffer. The resultant of bits is denoted by e_j.
- *Code block concatenation:* The input bit sequence is denoted by e^r_j where r is code block index. The output of code block concatenation and channel interleaving block is denoted by f_j. The code block concatenation concatenates the output of rate matching to different code blocks.
- *Multiplexing of data and control information:* In uplink, additionally, the control and data information is multiplexed in order to make the control information to be present on both slots in the subframe. Multiplexing ensures that control and data information is mapped to different modulation symbols.

11.7.3 OFDMA Downlink

OFDMA downlink implements downlink physical channels in order to convey information from higher layers. Scheduler in eNB allocates resource blocks to users for a predetermined amount of time. Downlink resource block, which is the smallest element of resource allocation, is illustrated in Fig. 11.24 where 12 subcarriers constitute a resource block with one downlink slot. LTE frames are 10 ms in duration

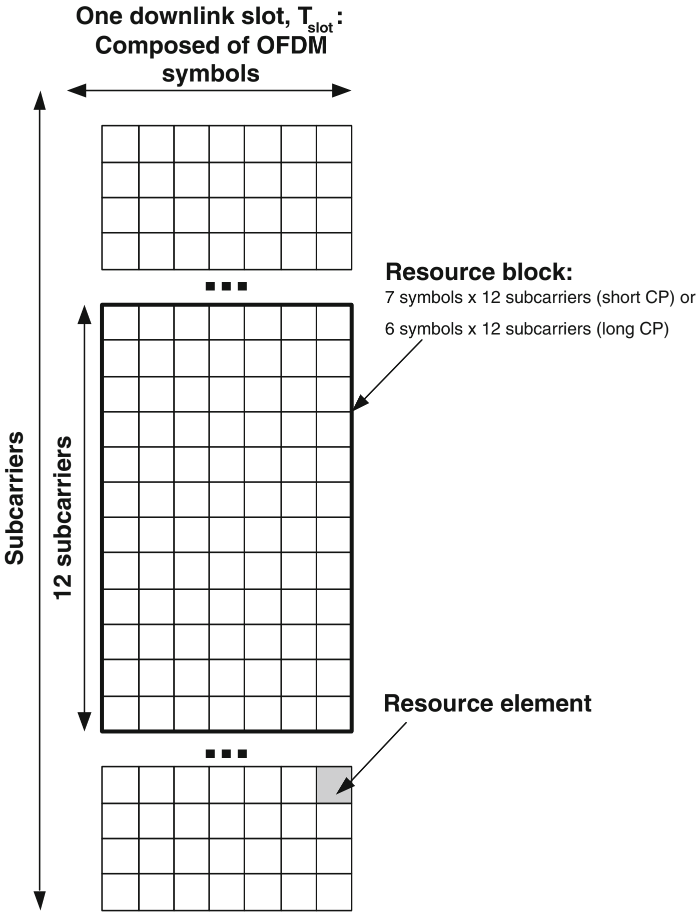

One downlink slot, T_{slot}: Composed of OFDM symbols
Subcarriers
12 subcarriers
Resource block:
7 symbols x 12 subcarriers (short CP) or
6 symbols x 12 subcarriers (long CP)
Resource element

Fig. 11.24 Downlink resource block

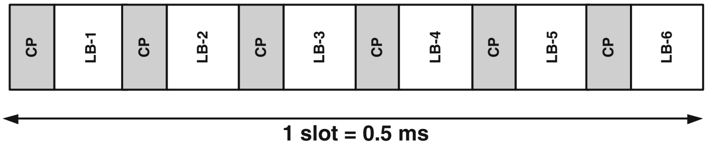


Fig. 11.25 Slot structure (0.5ms) for 3GPP LTE FDD downlink; *CP* cylic prefix, *LB* long blocks

and they are divided into ten subframes where each subframe is further divided into two slots, each of 0.5-ms duration. Slots consist of either 6 or 7 OFDM symbols as seen in Fig. 11.25 depending on whether the normal or extended cylic prefix is employed. Longer cylic prefix is desired to address longer fading that can be encountered in multicell broadcast service or very large cell deployments.

Table 11.3 Parameters for downlink transmission scheme for OFDMA

Transmission BW (MHz)	1.25	2.5	5	10	15	20
Subframe duration (ms)	0.5	0.5	0.5	0.5	0.5	0.5
Subcarrier spacing (KHz)	15	15	15	15	15	15
Sampling frequency (MHz)	1.92	3.84	7.68	15.36	23.04	30.72
FFT size	128	256	512	1,024	1536	2,048
No. of occupied subcarriers	76	151	301	601	901	1,201
Short/long CP	7/6	7/6	7/6	7/6	7/6	7/6
Short CP (s/samples)	(4.69/9)×6 (5.21/10)×1	(4.69/18)×6 (5.21/20)×1	(4.69/36)×6 (5.21/40)×1	(4.69/72)×6 (5.21/80)×1	(4.69/108)×6 (5.21/120)×1	(4.69/144)×6 (5.21/160)×1
Long CP (s/samples)	16.67/32	16.67/64	16.67/128	16.67/256	16.67/384	16.67/512
Resource block bandwidth (KHz)	180	180	180	180	180	180
No. of available RBs	6	12	25	50	75	100

Parameters for OFDMA downlink are shown in Table 11.3 for system bandwidths ranging from 1.25 to 20 MHz. The number of available subcarriers changes depending on the transmission bandwidth; however, OFDM downlink transmission scheme uses fixed subcarrier spacing regardless of transmission bandwidth. Basic downlink data-modulation schemes are QPSK, 16QAM, and 64QAM. Alternative enhanced modulation scheme is OFDM with pulse shaping (OFDM/OQAM), which does not require a guard interval.

Pilot symbols, which are the physical signals that terminate or originate in the PHY layer, are used to estimate the channel impulse response and for time and frequency synchronization. The orthogonal pilot sequence or a pseudo-random numerical sequence is used in order not to encounter an interference. A specific set out of 510 unique orthogonal sequences is assigned to each cell in order to distinguish a cell from others.

11.7.4 MIMO for OFDMA Downlink

The LTE PHY can employ multiple transceivers at the base station in order to enhance robustness and data rate.

MIMO 1 × 1, 2 × 2, 3 × 2, and 4 × 2 are supported either with single or multiuser MIMO. The maximum number of codewords is 2 regardless of the number of transmit antennas since there are at most two receive antennas. In addition, codebook-based precoding with a single precoding feedback and rank adaptation with single rank feedback are supported as well. Mapping of codewords is shown in Table 11.4; typical gains with MIMO are listed in Table 11.5.

Alamouti-based 2 × 1 MIMO is illustrated in Fig. 11.26. Downlink is considered as a double codeword system in which two parallel transmitters are used and combined at the later stage with a precoding matrix. OFDM mapper finally performs the modulation. However, uplink has only a single SC-FDMA transmitter. Pilot signals are transmitted at spaced subcarriers and channel impulse estimates for subcarriers that do not bear a pilot are computed with interpolation. When a pilot is transmitted in one antenna, the other antennas are set to idle to maintain the orthogonality.

Table 11.4 Single or double codeword assignment for MIMO

1 × 1	d0 →	Ant_1
2 × 2	d0 →	Ant_2
	d1 →	Ant_2
3 × 2	d0 →	Ant_1
	d1 →	Ant_2
	d1 →	Ant_3
4 × 2	d0 →	Ant_1
	d0 →	Ant_2
	d1 →	Ant_3
	d1 →	Ant_4

Table 11.5 OFDMA performance at 2.5 GHz with 10 MHz, TDD

MIMO	1×1	1×2	2×2	2×4	4×2	4×4
bps/Hz/sector	1.2	1.8	2.8	4.4	3.7	5.1

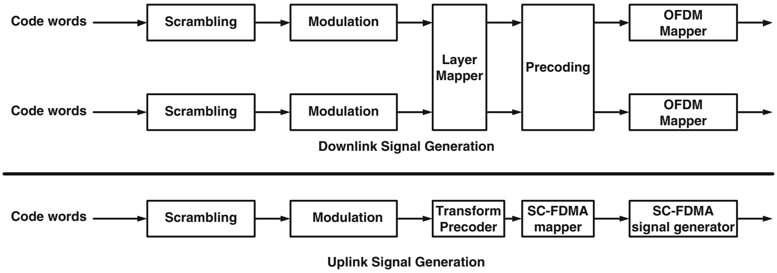


Fig. 11.26 Downlink and uplink signal generation for MIMO 2 × 1 LTE

11.7.5 SC-FDMA Uplink

SC-FDMA uplink is available for TDD and FDD modes. FDD is currently more popular and the uplink uses the same generic frame structure as in the downlink. Similarly, the same subcarrier spacing of 15KHz is used and parameters for uplink are given in Table 11.6.

Figure 11.27 shows the subframe structure for LTE: Long Blocks (LB) are used for control and data transmission and Short Blocks (SB) are used for pilot signals. In uplink, data are mapped onto a signal constellation that can be QPSK, 16QAM, or 64QAM with respect to channel quality. Modulated QPSK/QAM symbols are not directly used but sequentially fed into an FFT block to obtain the discrete frequency representation of the QPSK/QAM symbols. The FFT block is then mapped to an IFFT block to convert the sequence into time domain. This is to reduce the PAPR in the time domain since uncontrollable fluctuations are now restricted to the frequency domain because individual subcarrier amplitudes can actually vary more in the frequency domain as compared to OFDM. In brief, first FFT introduces high PAPR but second IFFT smoothes that over.

Table 11.6 Parameters for uplink transmission scheme for SC-FDMA

Transmission BW (MHz)	1.25	2.5	5	10	15	20
Subframe duration (ms)	0.5	0.5	0.5	0.5	0.5	0.5
LB/SB size (s)	66.67/33.33	66.67/33.33	66.67/33.33	66.67/33.33	66.67/33.33	66.67/33.33
LB/SB FFT size	128/64	256/128	512/256	1024/512	1536/768	2048/1024
Resource blocks	3	6	12	24	36	48
LB/SB # occupied subcarriers	75/38	150/75	300/150	600/300	900/450	1200/600
Short CP (s/samples)	3.65/7	3.91/15	4.04/31	4.1/63	4.12/95	4.13/127
Long CP (s/samples)	7.81/15	5.99/23	5.08/39	4.62/71	4.47/103	4.39/135

LB long blocks, *SB* short blocks

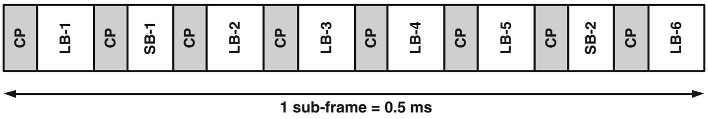


Fig. 11.27 Slot structure (0.5 ms) for 3GPP LTE FDD uplink; *CP* cylic prefix, *LB* long blocks (66.67 μs), *SB* short blocks (33.33 μs)

Physical layer procedures for uplink include link adaptation, uplink power control, and timing control. Also, control signaling for uplink consists of CQI, ACK/NAK, and scheduling request. CQI is to inform scheduler about current channel condition as well as feedback if MIMO is used. Uplink scheduling is done by eNB via the downlink CCPCH.

There are two physical signals: pilot symbols and random access preamble. Pilot symbols sent in the short blocks are used for coherent de/modulation and channel quality estimation. These reference symbols are constructed with CAZAC (Constant Amplitude Zero Autocorrelation) sequences. Zadoff-Chu[12] CAZAC polyphase sequences are defined as follows:

$$a_k = \begin{cases} \mathrm{e}^{-j2\pi \frac{r}{L}\left(\frac{k^2}{2}+qk\right)}, & k = 0,1,2,\cdots,L-1; \quad \text{for } L \text{ even} \\ \mathrm{e}^{-j2\pi \frac{r}{L}\left(\frac{k(k+1)}{2}+qk\right)}, & k = 0,1,2,\cdots,L-1; \quad \text{for } L \text{ odd,} \end{cases} \tag{11.1}$$

where r is an integer prime to L and q is any integer. Reference signal orthogonality can be achieved by nonoverlapping sequences in frequency, time, or code. Random access preamble, which is sent during cell search to initiate transmission, is based on Zadoff-Chu sequences as well. In FDD mode, there are 64 possible preamble sequences per cell site.

11.7.6 MIMO for SC-FDMA Uplink

E-UTRAN is being designed to support uplink peak data rate of 50 Mbps within 20-MHz uplink spectrum allocation (2.5 bps/Hz). Although in theory single antenna system can achieve 50 Mbps, at least 2 × 2 MIMO is considered to achieve the desired data rate.

Transmitter structure for MIMO SC-FDMA system is shown in Fig. 11.28 and receiver structure is shown in Fig. 11.29. One can see from the figure that transmitter can be configured for single or double codeword operation.

Performance comparison between single codeword and double codeword indicates that they show different performances in different SNR levels with transmit beamforming for 2 × 2 MIMO.[13] Performance of double codeword system shows higher throughput at lower SNR, and opposite is true at higher SNR. Double codeword system in the lower SNR region has one eigenmode that has higher SNR than the total system SNR. Consequently, that stream contributes some successful transmission while the lower stream generally does not. However, at high SNR, lower

[12] "The Zadoff-Chu CAZAC sequences show constant amplitude and flat frequency response. It has zero circular autocorrelation, and circular cross-correlation is low and constant provided that L is prime...".

[13] Simulation results presented in 3GPP document R1-063464 show that very high uplink spectral efficiency, about 2.8 bps/Hz, can be achieved using either method. And, double codeword can achieve 4 bps/Hz with 16QAM with different code rates on each stream, whereas single codeword must use a single code rate and different modulations.

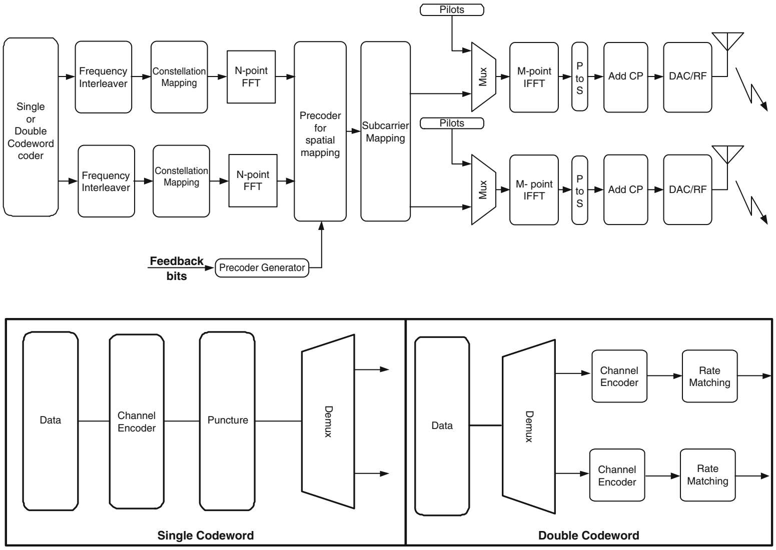


Fig. 11.28 MIMO SC-FDMA transmitter for single and double codewords

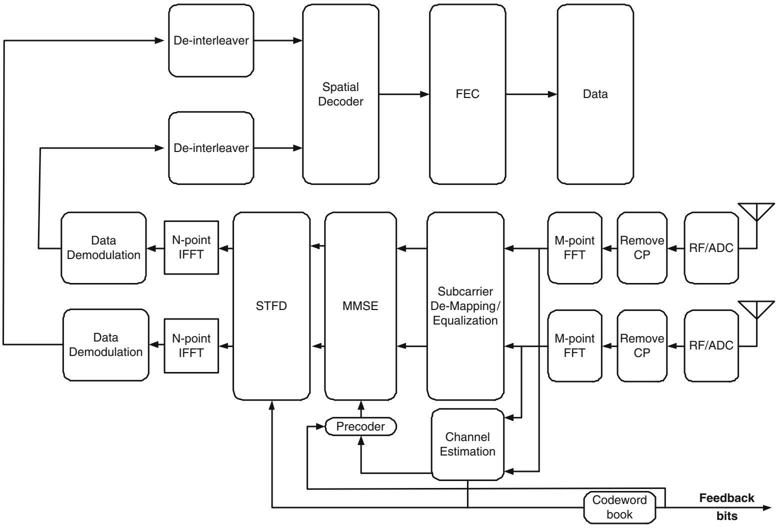


Fig. 11.29 MIMO SC-FDMA receiver for single codewords

stream reduces the total throughput because it still has a relatively dominant error rate. On the other hand, in single codeword system, the upper stream protects the lower stream due to the coding. As a result, lower error rate is achieved at higher SNRs.

11.8 Summary

Long-Term Evaluation (LTE) is the standardization work in 3GPP to offer a smooth evolutionary path for higher speeds and lower latency via OFDMA-based cellular architecture. System Architecture Evolution (SAE) being developed within 3GPP is an all IP-based system and intended for deployment with LTE. The key features of LTE are summarized here:

- LTE being specified in 3GPP Release 8 framework combines OFDMA downlink and SC-FDMA uplink.
- LTE supports scalable bandwidths ranging from 1.25 to 20 MHz, which enables LTE to operate in all 3GPP frequency bands in paired and unpaired spectrum allocations.
- With 20-MHz spectrum, theoretical rates are 300Mbps for downlink and 75 Mbps for uplink.
- LTE introduces E-UTRAN architecture comprising eNB as base station and UE as the mobile subscriber.
- SAE has separated the data and control plane in the access gateway and introduced MME for control functionality and SGW for data forwarding. Also PDN GW serves as a mobility anchor point.
- The capabilities of LTE will also evolve to fulfill the requirements of IMT-Advanced with LTE *Advanced*.

References

1. 3GPP TS 23.402, "Architecture enhancements for non-3GPP accesses." http://www.3gpp.org.
2. ITU T Recommendation I.130, "Method for the characterization of telecommunication services supported by an ISDN and network capabilities of an ISDN." http://www.itu.int.
3. ITU T Recommendation Q.65, "The unified functional methodology for the characterization of services and network capabilities." http://www.itu.int.
4. 3GPP TS 36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2." http://www.3gpp.org.
5. 3GPP TS 23.203, "Policy and charging control architecture." http://www.3gpp.org.
6. 3GPP TS 23.060, "General Packet Radio Service (GPRS); Service description; Stage 2." http://www.3gpp.org.
7. 3GPP TS 43.129, "Packet-switched handover for GERAN A/Gb mode; Stage 2." http://www.3gpp.org.
8. 3GPP TS 23.003, "Numbering, addressing and identification." http://www.3gpp.org.
9. 3GPP TS 23.122, "Non-Access-Stratum (NAS) functions related to Mobile Station in idle mode." http://www.3gpp.org.
10. 3GPP TS 43.022, "Functions related to MS in idle mode and group receive mode." http://www.3gpp.org.
11. 3GPP TS 25.304, "UE procedures in idle mode and procedures for cell re-selection in connected mode." http://www.3gpp.org.

12. 3GPP TS 23.246, “Multimedia Broadcast/Multicast Service (MBMS); Architecture and functional description.” http://www.3gpp.org.
13. 3GPP TS 29.060, “GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface.” http://www.3gpp.org.
14. 3GPP TS 43.051, “GERAN Overall description – Stage 2.” http://www.3gpp.org.
15. 3GPP TS 25.401, “UTRAN overall description.” http://www.3gpp.org.
16. IETF RFC 1034 (1987), “Domain names – concepts and facilities (STD 13).” http://www.ietf.org.
17. IETF RFC 4862, “IPv6 Stateless Address Autoconfiguration.” http://www.ietf.org.
18. IETF RFC 2131, “Dynamic Host Configuration Protocol.” http://www.ietf.org.
19. IETF RFC 3736, “Stateless Dynamic Host Configuration Protocol (DHCP) Service for IPv6.” http://www.ietf.org.
20. IETF RFC 3633, “IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6.” http://www.ietf.org.
21. 3GPP TS 25.413, “UTRAN Iu interface Radio Access Network Application Part (RANAP) signalling.” http://www.3gpp.org.
22. 3GPP TS 44.064, “Mobile Station – Serving GPRS Support Node (MS-SGSN); Logical Link Control (LLC) Layer Specification.” http://www.3gpp.org.
23. 3GPP TS 23.251, “Network Sharing; Architecture and functional description.” http://www.3gpp.org.
24. IETF RFC 4039, “Rapid Commit Option for the Dynamic Host Configuration Protocol version 4 (DHCPv4).” http://www.ietf.org.
25. IETF RFC 768, “User Datagram Protocol.” http://www.ietf.org.
26. 3GPP TS 23.221, “Architectural requirements.” http://www.3gpp.org.
27. 3GPP TS 23.008, “Organization of subscriber data.” http://www.3gpp.org.
28. 3GPP TS 23.078, “Customized Applications for Mobile network Enhanced Logic (CAMEL) Phase X; Stage 2.” http://www.3gpp.org.
29. 3GPP TS 23.236, “Intra-domain connection of Radio Access Network (RAN) nodes to multiple Core Network (CN) nodes.” http://www.3gpp.org.
30. IETF RFC 3588, “Diameter Base Protocol.” http://www.ietf.org.
31. IETF RFC 4861, “Neighbor Discovery for IP Version 6 (IPv6).” http://www.ietf.org.
32. 3GPP TS 25.331, “Radio Resource Control (RRC); Protocol Specification.” http://www.3gpp.org.
33. 3GPP TS 36.304, “Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode.” http://www.3gpp.org.
34. IETF RFC 2960, “Stream Control Transmission Protocol.” http://www.ietf.org.
35. 3GPP TS 36.413, “Evolved Universal Terrestrial Access Network (E-UTRAN); S1 Application Protocol (S1AP).” http://www.3gpp.org.
36. 3GPP TS 36.331, “Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification.” http://www.3gpp.org.
37. 3GPP TS 29.061, “Interworking between the Public Land Mobile Network (PLMN) supporting packet based services and Packet Data Networks (PDN).” http://www.3gpp.org.
38. IETF RFC 3041, “Privacy Extensions for Stateless Address Autoconfiguration in IPv6.”
39. 3GPP TS 33.102, “3G Security; Security architecture.” http://www.3gpp.org.
40. 3GPP TS 33.401, “3GPP System Architecture Evolution: Security Architecture.” http://www.3gpp.org.
41. 3GPP TS 36.300, “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2 (Release 8).” http://www.3gpp.org.
42. 3GPP TR 25.913, “Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN).” http://www.3gpp.org.
43. 3GPP TS 36.201, “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; General description.” http://www.3gpp.org.
44. 3GPP TS 36.211,“Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation.” http://www.3gpp.org.

45. 3GPP TS 36.212, "Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding." http://www.3gpp.org.
46. 3GPP TS 36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures." http://www.3gpp.org.
47. 3GPP TS 36.214, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements." http://www.3gpp.org.
48. 3GPP TS 36.302, "Evolved Universal Terrestrial Radio Access (E-UTRA); Services provided by the physical layer." http://www.3gpp.org.
49. 3GPP TS 36.304, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) procedures in idle mode." http://www.3gpp.org.
50. 3GPP TS 36.306, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio access capabilities." http://www.3gpp.org.
51. 3GPP TS 36.321, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Acces Control (MAC) protocol specification." http://www.3gpp.org.
52. 3GPP TS 36.322, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Link Control (RLC) protocol specification." http://www.3gpp.org.
53. 3GPP TS 36.323, "Evolved Universal Terrestrial Radio Access (E-UTRA); Packet Data Convergence Protocol (PDCP) specification." http://www.3gpp.org.
54. 3GPP TS 36.331, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC) protocol specification." http://www.3gpp.org.
55. 3GPP TS 23.401, "Technical Specification Group Services and System Aspects; GPRS enhancements for E-UTRAN access." http://www.3gpp.org.
56. 3GPP TR 24.801, "3GPP System Architecture Evolution (SAE); CT WG1 aspects." http://www.3gpp.org.
57. 3GPP TS 23.402, "3GPP System Architecture Evolution: Architecture Enhancements for non-3GPP accesses." http://www.3gpp.org.
58. 3GPP TR 25.913 V7.2.0 (2005-06), "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)." http://www.3gpp.org.
59. R1-050665, NTT DoCoMo, "Throughput Evaluations Using MIMO Multiplexing in Evolved UTRA Uplink," 3GPP RAN WG1 LTE, June 2005. http://www.3gpp.org.
60. R1-060437, NTT DoCoMo, NEC, Sharp, Toshiba Corporation, "Basic Schemes in Uplink MIMO Channel Transmissions," 3GPP TSG RAN WG1 Meeting #44, February 2006. http://www.3gpp.org.
61. R1-061481, InterDigital, "User Throughput and Spectrum Efficiency for E-UTRA," 3GPP RAN1 LTE, May 2006. http://www.3gpp.org.
62. R1-060365, InterDigital, "Extension of Uplink MIMO SC-FDMA with Preliminary Simulation Results," 3GPP RAN1 LTE, February 2006. http://www.3gpp.org.
63. R1-062162, InterDigital, "Uplink SDMA-MU-MIMO using Precoding for E-UTRA," 3GPP TSG-RAN WG1 Meeting #46, August 28 – September 1, 2006. http://www.3gpp.org.
64. R1-062160, InterDigital, "Uplink MIMO Precoding Using Differential Feedback," 3GPP TSG-RAN WG1 Meeting #46, August 28 – September 1, 2006. http://www.3gpp.org.
65. R1-063466, InterDigital, "Uplink MIMO Precoding Feedback," 3GPP TSG-RAN WG1 Meeting #47, Riga, Latvia, November 6 – 10, 2006. http://www.3gpp.org.
66. R1-051344, Samsung, "Downlink Pilot and Control Channel Structure for E-UTRA," 3GPP TSG RAN WG1 Meeting #43, Seoul, Korea, November 7–11, 2005. http://www.3gpp.org.
67. 3GPP TD RP-040461, "Proposed Study Item on Evolved UTRA and UTRAN." http://www.3gpp.org.
68. Ekström, H., Furuskär, A., Karlsson, J., Meyer, M., Parkvall, S., Torsner, J., Wahlqvist, M., "Technical Solutions for the 3G Long-Term Evolution," *IEEE Commun. Mag.*, vol. 44, no. 3, pp. 38–45, 2006.
69. Ojanpera, T and Prasad, R., "An Overview of Air Interface Multiple Access for IMT2000/UMTS," *IEEE Commun. Mag.*, pp. 82–95, 1998.
70. Ojanpera, T., "Overview of Research Activities for Third Generation Mobile Communication: Wireless Communications TDMA versus CDMA," Kluwer Academic, pp. 437–439, 1998.
71. IETF RFC 2960 (10/2000), "Stream Control Transmission Protocol." http://www.ietf.org.

Chapter 12
Ultra Mobile Broadband of 3GPP2

12.1 Introduction

Ultra Mobile Broadband (UMB)[1] of Third Generation Partnership Project 2 (3GPP2) considers OFDMA and all-IP-based network to fulfill the ITU vision for next generation services in order to enable the convergence of IP-based voice, broadband data, multimedia, information technology, entertainment, and consumer electronic services. The UMB is being designed to offer economical support for a large variety of services that require extremely low latency, low jitter, and high spectral efficiencies to address a large cross section of advanced mobile broadband services.

3GPP2's convergence to UMB follows an evolutionary path in the family of CDMA2000 standards with recent Flash-OFDM (see Sect. 5.7) input as seen in Fig. 12.1. UMB is therefore a candidate for existing or new spectrum allocations with its scalable bandwidth feature up to 20 MHz. UMB enhances the performance and capabilities by combining the best aspects of CDMA, TDM, OFDM, and OFDMA with MIMO and SDMA advanced antenna techniques.

With the ability to support peak download speeds as high as 280 Mbps in a mobile environment, Ultra Mobile Broadband (UMB), standardized in 2007,[2] offers the following features in the delivery of next generation mobile broadband services:

- Peak download and upload rates of 288 Mbps and 75 Mbps, respectively in 20 MHz bandwidth.
- Connectivity for voice and data in all environments; fixed, pedestrian, and full-mobile up to 300 km/hr.

[1] The CDMA Development Group (CDG) has selected "Ultra Mobile Broadband" as the brand name to describe the advanced technologies and services that will be supported by the CDMA2000 1xEV-DO Revision C (Rev. C) standard.

[2] The CDMA Development Group (CDG) (www.cdg.org) and the Third Generation Partnership Project 2 (3GPP2) (www.3GPP2.org) announced in September 2007 the publication of the Ultra Mobile Broadband^{TM} (UMBTM) air interface specification 3GPP2 C.S0084-0 v2.0 - Set of standards that constitute UMB is listed here:

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_12, © Springer Science+Business Media LLC 2009

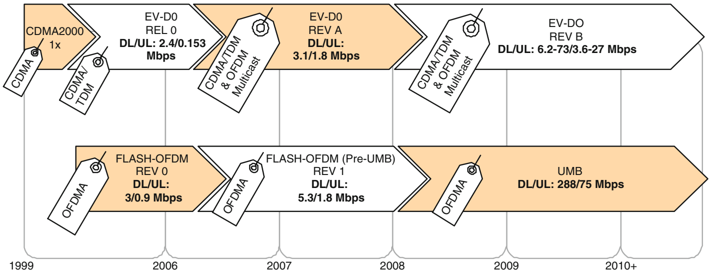


Fig. 12.1 UMB evolution path: EV-DO REV B peak rates are scalable with number of carriers – standard supports up to 15 carriers. Upper range is with 64QAM where 1 RF Carrier has 4.9 Mbps peak. UMB peak rate is with 20 MHz bandwidth and 4 × 4 MIMO in FDD mode (source: CDG)

- Low latency of 14.3 ms for delay-sensitive applications such as VoIP, push-to-talk, etc.
- Up to 1,000 simultaneous mobile VoIP sessions within a sector with 20 MHz bandwidth.
- Seamless handover with wide area coverage or hot zone coverage.
- Converged access network (CAN), which is an IP-based Radio Access Network (RAN) of 3GPP2.

- C.S0084-00 – Overview
- C.S0084-001 – Physical Layer
- C.S0084-002 – MAC Layer
- C.S0084-003 – Radio Link Layer
- C.S0084-004 – Application Layer
- C.S0084-005 – Security Functions
- C.S0084-006 – Connection Control Plane
- C.S0084-007 – Session Control Plane
- C.S0084-008 – Route Control Plane
- C.S0084-009 – Broadcast-Multicast Upper Layers

TSG-X group is working to standardize networking layer within X.S0054-000-X-n series (Interoperability Specification is A.S0020). The first version was published in December 2007. These documents define an architecture model and set of specifications for a Converged Access Network (CAN) to support multiple technologies including UMB air interface. The set of standards that constitute CAN include the following:

- X.S0054-000-0 v1.0 – CAN Wireless IP Network Overview and List of Parts
- X.S0054-100-0 v1.0 – Basic IP Services for Converged Access Network Specifications
- X.S0054-102-0 v1.0 – Multiple Authentication and Legacy Authentication Support for Converged Access Network
- X.S0054-110-0 v1.0 – MIPv4 Specification in Converged Access Network
- X.S0054-210-0 v1.0 – CMIP-Based Inter-AGW Handoff
- X.S0054-220-0 v1.0 – Network PMIP Support
- X.S0054-300-0 v1.0 – QoS Support for Converged Access Network Specification.
- X.S0054-400-0 v1.0 – Converged Access Network Accounting Specification
- X.S0054-910-0 v1.0 – CAN Data Dictionary

Published TSG-C and TSG-X specifications can be found at `http://www.3gpp2.org/`.

- Support for flexible bandwidth between 1.25 and 20 MHz using channel bandwidth allocations around 150 KHz within the 450, 700, 850, 1700, 1900, 1700/2100 (AWS), 1900/2100 (IMT), and 2500 MHz (3G extension) spectrum bands.
- Support for multicarrier deployments where two bandwidths are concatenated to support wider band. Low complexity terminals and wideband terminals can be supported concurrently.
- Support for FDD/TDD modes.
- Support for full frequency reuse with no frequency planning.
- Support for MIMO and SDMA.
- Support for convergence toward ubiquitous and universal access. It supports intertechnology handoffs with WiMAX, 3GPP and seamless operation with existing CDMA2000 1X and 1xEV-DO systems.

In this chapter, we introduce UMB technology in detail along with its all-IP-based network architecture, termed CAN. We first describe the UMB reference model followed by CAN. Then, we talk about UMB air interface and focus on PHY and MAC features including resource allocation and MIMO & SDMA support.

12.2 Reference Model

UMB reference model consists of Access Terminals (ATs) and Access Networks (ANs) as seen in Fig. 12.2. An Access Terminal has a radio interface to communicate with the Access Networks.

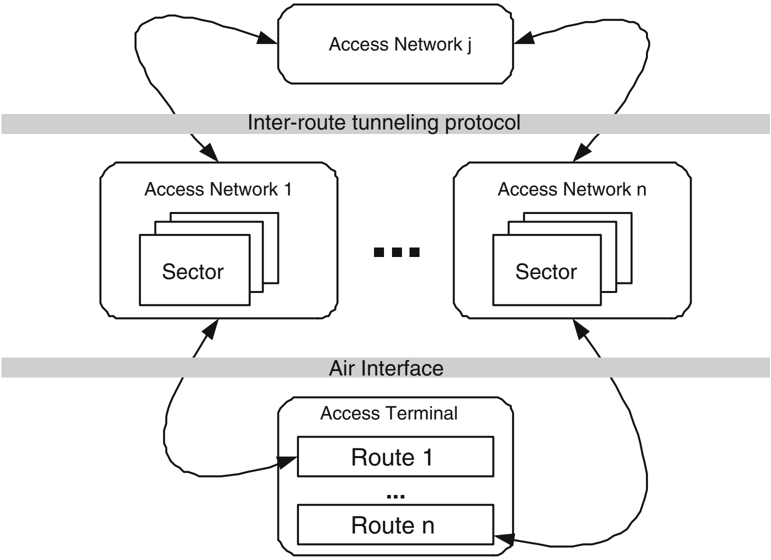


Fig. 12.2 UMB architecture reference model

An AN is a network entity that contains an Access Network Route Instance (ANRI) for the purpose of logically communicating with the AT. An AT maintains an *InUse* instance (aka Route) that is associated with each AN it is in communication with. ANs may talk to each other through Route Protocol using Inter-Route Tunneling Protocol. Thus, AT may initiate multiroute with different ANs at a given time. We discuss multiroute functionality at later section in this chapter. UMB utilizes CAN as the Access Network.

12.3 CAN: Converged Access Network

Converged Access Network (CAN) defined to support UMB within 3GPP2 is shown in Fig. 12.3. CAN architecture is composed of Access Gateways (AGWs), evolved Base Stations (eBSs), and Session Reference Network Controllers (SRNCs) in addition to Access Terminals (ATs), Home Agent (HA), AAA server, etc. UMB employs

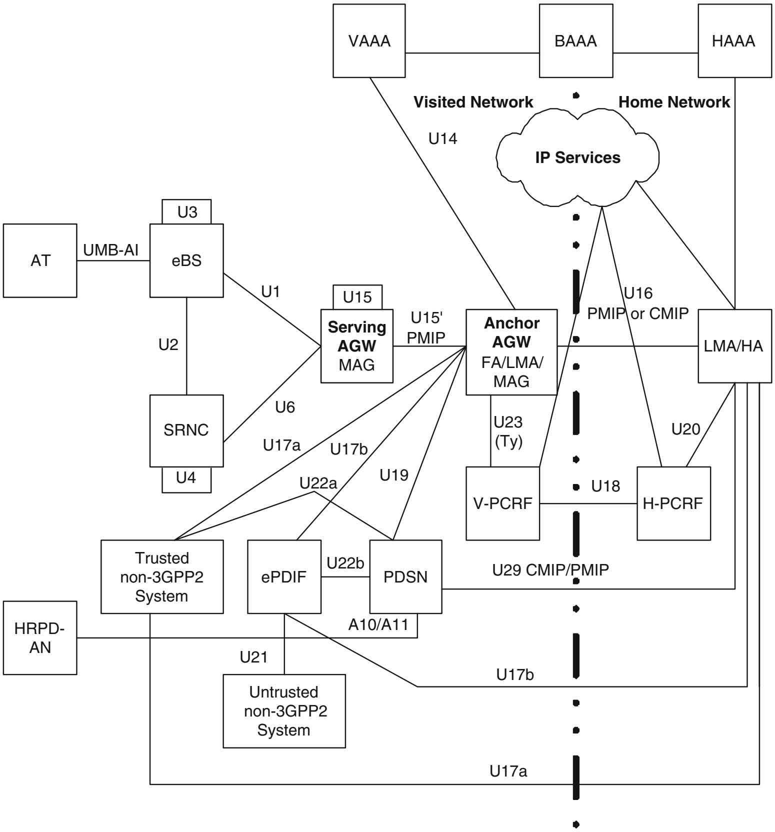


Fig. 12.3 CAN for roaming scenario with split AGW (© 3GPP2)

IETF-based protocols between most of the entities in order to simplify the IOT (Interoperability Testing) between vendors, and lower the development cost and time. For instance, SRNC and AGW communicate using AAA protocol for EAP proxy whereas AGW and AAA communicate via AAA protocol for EAP proxy and accounting. For mobility, CMIP is used for inter-AGW handover but there is a support for simple IP and PMIP. 3GPP2 standardizes its own protocols between SRNCs, eBSs, and SRNC and eBS.

12.3.1 AGW: Access Gateway

Access Gateway (AGW) provides user's point of IP connectivity to the network. AGW is first-hop IP router to the packet data network, which tunnels the IP packets for user data to eBS as seen in Fig. 12.4. AGW holds the AAA client for accounting broker, AAA proxy for AAA client in SRNC, and foreign agent if applicable. In split-mode, there is a Serving AGW (srv-AGW) and an Anchor AGW (anc-AGW). AGW may hold Proxy Mobile Agent (PMA) in IPv4 or Mobile Access Gateway (MAG) in IPv6 to manage the mobility-related signaling for a mobile node that is attached to its access link. Anc-AGW may also hold Local Mobility Anchor (LMA), which is the home agent for the mobile node in the PMIPv6. It is the anchor point for the AT's home prefix and maintains the reachability state.

AGW provides QoS control of the traffic and flow-based packet counting functionalities to support online and offline charging. The AGW has Policy Enforcement Point (PEP) for QoS policy and local resource-based policy and Traffic Plane Function (TPF) to apply the charging rules statistically provisioned or dynamically provided by a Charging Rules Function for the purposes of offline and/or online charging. AGW is Ty termination point to PCRF and employs QoS enforcement point of DSCP marking based on Ty policies.

In addition to these, AGW also provides intrusion detection of access points and DHCP functionalities for IP address assignment to AT.

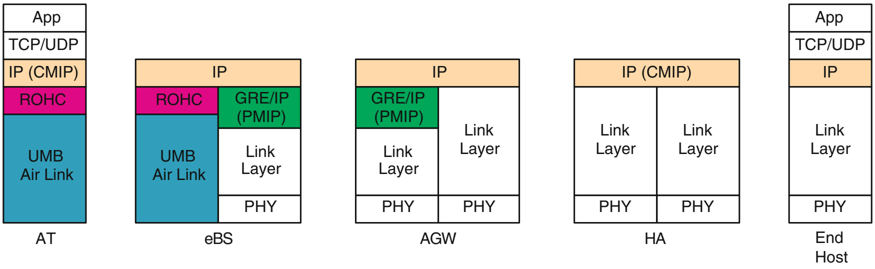


Fig. 12.4 UMB protocol stack (© 3GPP2)

12.3.2 SRNC: Session Reference Network Controller

Session Reference Network Controller (SRNC) performs control signaling functions for the CAN. SRNC has reduced user plane functionalities compared with RNC since there is no QoS and no RLP (Radio Link Protocol), and RNC functionalities in 1xEV-DO are distributed among AGW, SRNC, and eBS in UMB. SRNC holds the data base for sessions and performs paging and location management together with idle state management. It also holds the authenticator for EAP-based access. SRNC can either be a separate entity apart from eBS and AGW or it can be coupled with eBS. The SRNC holds *Session Anchor*, which keeps ANRI for each AT that is supporting as seen in Fig. 12.5.

12.3.3 eBS: Evolved Base Station

The evolved Base Station (eBS) holds the UMB air interface stack and networking stack. eBS supports over-the-air (OTA) bearer communication with the AT through following states:

- FLSE: *Forward Link Serving eBS* is the anchor point for forward link (FL) packets. Packets coming from AGW are forwarded to AT through FLSE eBS.
- RLSE: *Reverse Link Serving eBS* is the anchor point for reverse link (RL) packets. Packets sent by AT are received by RLSE eBS that is forwarded to DAP.

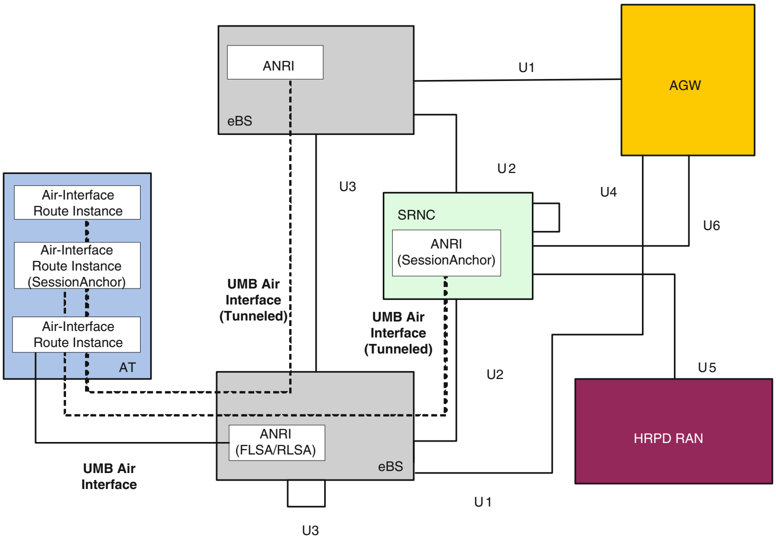


Fig. 12.5 ANRI (© 3GPP2)

- DAP: *Data Attachment Point* is the point-of-contact of eBS to the AGW. AGW sends the data and control signaling-related information of AT to the DAP. Reverse link information is sent to DAP to be forwarded to AGW by AT or by the RLSE eBS. The DAP relocation can happen with respect to several performance criteria. One such criterion is moving DAP when DAP eBS is no longer in the active set of AT.

eBS is also responsible for OTA QoS scheduling and Admission Control together with a packet classifier, an accounting client, header compression client (RoHC), and security functionalities such as user data ciphering, encryption, and decryption. eBS may contain functionalities of SRNC depending on the application.

12.3.4 Other Entities

Besides Access Terminals (AT) that implement the UMB air interface described in the next section, other entities depicted in Fig. 12.3 of the CAN are explained in the following:

- *AAA:* AAA provides authentication, authorization, and accounting functions. The AAA has three roles: Visited (V), Broker (B), and Home (H) AAA.
- *HA:* Home Agent (HA) provides mobility solution to the AT in the 3GPP2. HA keeps AT's reachability state. Depending on the configuration home agent may have the Local Mobility Agent (LMA).
- *HRPD-AN:* High-Rate Packet Data Access Network (HRPD-AN) is a node in the HRPD legacy packet data network.
- *ePDIF:* Evolved Packet Data Interworking Function (ePDIF) facilitates the connection between 3GPP2 and an untrusted non-3GPP2 network such as WiFi. Also fixed mobile convergence is through Packet Data Interworking Function (PDIF). PDIF is responsible to facilitate security, access, authentication, and policy enforcement along with other respective entities in the network.
- *PDSN:* Packet Data Serving Node (PDSN) facilitates the integration to existing CDMA2000 network. PDSN provides access to the Internet, intranets, and application servers for mobile stations in a CDMA2000 Radio Access Network (RAN). Acting as an access gateway, PDSN provides Simple IP and Mobile IP access, Foreign Agent support, and packet transport for virtual private networking. It acts as a client for AAA servers and provides mobile stations with a gateway to the IP network.
- *PCRF:* Policy and Charging Rules Function (PCRF) detects a packet belonging to a service data flow to provide policy control and applicable charging along with IMS (see last chapter).

12.3.5 Reference Points

CAN entities are connected through interfaces that are standardized by the reference points. The reference points, most of which seen in Fig. 12.3, are explained here:

- *UMB-AI* is the air interface and carries information between AT and eBS.
- *U1* is between eBS and AGW for control and bearer information.
- *U2* is between eBS and SRNC to carry control information.
- *U3* is between eBSs to carry control and bearer information.
- *U4* is between SRNCs to carry control information.
- *U5* is between SRNC and HRPD RAN to carry control information for handover.
- *U6* is between SRNC and AGW to carry control information and data during paging.
- *U14* is between AGW and AAA via VAAA to carry signaling AAA information.
- *U15* is between srv-AGWs to carry PMIP information.
- *U15′* is between srv-AGW and anc-AGW in split-mode to carry policy information.
- *U16* is between AGW and HA to carry control and bearer information.
- *U17* is between 3GPP2 network and a non-3GPP network to carry control and bearer information. U17a is for trusted network such as WiMAX or 3GPP SAE/LTE and U17b is for untrusted network such as WiFi, FMC, etc.
- *U18* is between Visitor-PCRF and Home-PCRF to carry policy and charging rules.
- *U19* is between AGW and non-3GPP2 system to carry information for fast/seamless handoff between CAN and another system. In single AGW mode, U19a is to trusted non-3GPP2 system, U19b is to ePDIF, U19c is to PDSN, in split-mode, U19 is to PDSN from anc-AGW.
- *U20* is between HA and PCRF to convey QoS policy and charging information to HA.
- *U21* is between ePDIF and untrusted non-3GPP2 Access Network.
- *U22* is between PDSN to un/trusted non-3GPP2 network. U22a is to trusted non-3GPP2 network and U22b is to untrusted non-3GPP2 via ePDIF.
- *U23* (aka Ty) is between PCRF and anc-AGW/AGW to carry control information.
- *U24* is between LMA and another system to carry control and bearer information with PMIP. U24a is to trusted non-3GPP2 network, U24b is to untrusted non-3GPP2 via ePDIF, and U24c is to PDSN.
- *U25* is between HA and another system to carry control and bearer information with CMIP. U25a is to trusted non-3GPP2 network and U25b is to untrusted non-3GPP2 via ePDIF.
- *U26* is between LMA and AGW to carry control and bearer information with PMIP to support handover between AGWs.
- *U27* is between HA and AGW to carry control and bearer information with CMIP to support handover between AGWs.
- *U28* is between AGWs to support fast inter-AGW handoff.

- *U29* is between HA and PDSN to carry control and bearer information. In split-mode, it carries CMIP and PMIP, otherwise only CMIP.
- *A10/11* is between PDSN and HRPD AN to carry control (A11) and bearer (A10) information.

12.4 Mobility Support

The mobility scenarios for UMB consider three types of mobility: within AGW domain, across AGW domain, and between different technologies. We first talk about *multiroute* feature of UMB that is introduced to provide flexible and seamless handover experience and then explain the details of these mobility models.

12.4.1 Multiroute

Multiroute functionality enables AT to establish more than one *route* with different eBSs in the active set. Route is defined as a protocol stack, which is instantiated per eBS. Maximum 6 routes are supported when active; however, in idle mode, only one route with SRNC is supported. This allows AT to maintain FL with one eBS and RL with another eBS, simultaneously.

Multiroute aims to provide fast switching between eBSs with low associated overhead. AT can communicate to eBS before any handoff and has the advantage of selecting the best serving eBS for optimized OTA performance with minimum overhead of switching.

Multiroute is also designed to integrate the future eBS that has different capabilities or revisions with no modification since it simplifies the inter-eBS handover by removing the need to transfer connected state information. This allows eBS and SRNC to be independently and gradually upgraded.

12.4.2 Inter-eBS handover

AT can change the eBS through Layer 2 handover. UMB offers one additional form of flexibility in handover by assigning a handover functionality to FL and RL separately. Thus, eBS for new FL and new RL can be selected independently by moving FLSE or RLSE to new eBS, respectively. For instance, new RL can be established with a target eBS while preserving the old FL with the current eBS. AT decides for FLSE or RLSE and notifies eBS. New FLSE or RLSE sends *IPT-Notification* messages to DAP. If a FLSE change is requested, DAP forwards packets to new FLSE with IP tunnel. If a RLSE change is requested, new RLSE may send packets either to DAP or to AGW depending on the configuration.

DAP transfer is established after FLSE. Before DAP transfer, source DAP continues to tunnel data to target eBS until L3 handover occurs. Target DAP eBS performs PMIP update with AGW and then notifies AT, the serving DAP, and SRNC. DAP transfer can be initiated either from AT or through AN. Multiroute feature may delay the tunnel binding with AGW depending on the performance criteria.

SRNC transfer happens when AT crosses a new paging region. SRNC transfer involves moving session storage, paging, and idle state management to new SRNC. Target SRNC communicates to serving SRNC in order to send UATI update to serving SRNC. Then, target SRNC notifies the AT with new UATI signal.

12.4.3 Inter-AGW handover

Inter-AGW handover occurs when an eBS in the active session cannot communicate to the serving AGW. Inter-AGW handover does not require an interface between AGWs and there is no data interruption during inter-AGW handover. AT is served by one DAP, and L2 tunneling is established between eBSs under different AGWs. However, two PMIP tunnels are maintained for a short period.

12.4.4 Intersystem handover

Converged Access Network enables seamless handover by establishing tunnels between UMB and other access technologies. Figure 12.3 shows the reference point mapping to existing 3GPP2 networks and un/trusted non-3GPP2 networks. AGW is in communication with PDSN for handover to existing 3GPP2 EV-DO networks. Handover to WiFi, DSL is through ePDIF and AGW. On the other hand, if non-3GPP2 network is trusted, such as WiMAX or LTE, it can directly communicate with AGW.

12.5 UMB Air Interface Protocol Architecture

Now, we start talking about UMB air interface. UMB air interface is between AT and eBS as seen in Fig. 12.3. Layers and interaction of UMB air interface with end-to-end session are illustrated in Fig. 12.4. UMB air interface, depicted in Fig. 12.6, adopts OFDMA for forward and reverse link and OFDMA together with CDMA for RL control channel information. In this section, we summarize the UMB air interface layers and starting from next section we focus on physical and MAC layers.

- *Physical layer* provides channel structure, frequency, power output, modulation, and encoding specifications for the FL and RL.

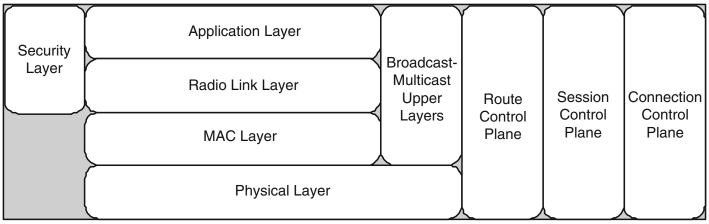


Fig. 12.6 UMB air interface

- *MAC layer* defines the procedures used to receive and transmit over the PHY layer:
  - *Packet Consolidation Protocol* transmits prioritization and packet encapsulation for upper layer packets.
  - *Superframe Preamble MAC Protocol* provides procedures about the transmission and reception of the superframe preamble.
  - *Access Channel MAC Protocol* provides the transmission and reception procedures for Access Channel.
  - *Forward Link Control Segment MAC Protocol* provides procedures for Forward Control Channel.
  - *Forward Traffic Channel MAC Protocol* provides transmission and reception procedures for the Forward Data Channel.
  - *Reverse Traffic Channel MAC Protocol* provides the transmission and reception procedures for the Reverse Control Channel.
  - *Reverse Control Channel MAC Protocol* provides the transmission and reception procedures for the Reverse Data Channel.

- *Radio Link Layer* provides services such as reliable and in-sequence delivery of application layer packets together with multiplexing of application layer packets and QoS negotiation:
  - *QoS Management Protocol* provides QoS to application layer packets with appropriate flow and filter specifications.
  - *Radio Link Protocol* provides fragmentation and reassembly, retransmission, and duplicate detection for upper layer packets.
  - *Stream Protocol* provides mapping of upper layer fragments to particular streams.
  - *Route Protocol* provides routing of Stream Protocol packets over the air link.

- *Application Layer* provides multiple applications:
  - *Signaling Protocol* provides message transmission services for signaling messages.
  - *Interroute Tunneling Protocol* provides transport of packets from other routes.
  - *Other Application Layer Protocols* such as IP, EAP, and RoHC may create payload to be carried over the UMB air interface.

- *Security Functions* include several protocols:
  - *Key Exchange Protocol* provides the means to generate security keys between AT and AN.
  - *Ciphering Protocol* provides the procedures for ciphering traffic.
  - *Message Integrity Protocol* provides the procedures for integrity protection of signaling messages.
- *Connection Control Plane* provides air link connection establishment and maintenance services:
  - *Air Link Management Protocol* provides management of state machines during a connection.
  - *Initialization State Protocol* provides the procedures for network entry.
  - *Idle State Protocol* provides the procedures when AT is in idle mode.
  - *Connected State Protocol* provides procedures that an AT and an AN follow when a connection is open.
  - *Active Set Management Protocol* provides procedures to maintain air link.
  - *Overhead Messages Protocol* provides broadcast messages that are used in connection layer protocols.
- *Session Control Plane* provides negotiation and configuration of the protocols used in the session.
- *Route Control Plane* performs creation, maintenance, and deletion of routes.
- *Broadcast-Multicast Service* (BCMCS) procedures are to offer multicast and broadcast services:
  - *Broadcast Interroute Tunneling Protocol* provides transport for packets from other routes.
  - *Broadcast Packet Consolidation Protocol* provides framing of BCMCS content packets and multiplexing of packets to be sent over BCMCS channel.
  - *Broadcast Security Protocol* provides ciphering of content.
  - *Broadcast MAC Protocol* provides the procedures of transmission by AN and reception by the AT.
  - *Broadcast Control Protocol* provides control procedures such as registration.

12.6 UMB Physical and MAC Layers

We talk about underlying principles in frame construction including permutation characteristic for resource allocation to exploit frequency or multiuser diversity. Then, we talk about the supported coding and modulation schemes and OFDM structure with parameters. First, we introduce FL and RL channels since the PHY and MAC are defined in terms of channels as can be seen in the next sections.

12.6.1 Forward and Reverse Link Channels

Transmissions are performed over many different channels. Each channel is responsible to carry certain information or data. Channels introduced below facilitate quality and continuity of the transmission in the air link. We also give the encoding methods of each channel. Channels for FL are as follows:

- *Forward Preamble Pilot Channel* is used by an AN to aid acquiring the system. No CRC and FEC is applied.
- *Forward Other Sector Interference Channel* is used for the transmission of the FL preamble, which consists of two OFDM symbols to help initial acquisition process. Also, these symbols carry information about other sector interference received by SFP MAC Protocol. No CRC and FEC is applied.
- *Forward Primary Broadcast Control Channel* is used for preamble transmission to inform about deployment-specific parameters such as cylic prefix duration, number of guard carriers, in addition to superframe index. CRC-12 and Rate 1/3 Convolutional encoding is applied with channel interleaver and sequence repetition.
- *Forward Secondary Broadcast Control Channel* is used for preamble transmission in order to carry information for AT to demodulate the data from PHY frames. Information includes hopping patterns, pilot structures, control channel structures, effective antennas, multiplexing modes, etc. CRC-12 and Rate 1/3 Convolutional encoding is applied along with channel interleaver and sequence repetition.
- *Forward Acquisition Channel* is sent on a preamble consisting of one OFDM symbol to help the initial acquisition process. No CRC and FEC is applied.
- *Forward Beacon Pilot Channel* is used to indicate the presence of AN. No CRC and FEC is applied.
- *Forward Quick Paging Channel* is used on the FL link preamble to help AT identify the paging signal. CRC-12 and Rate 1/3 Convolutional encoding is applied with channel interleaver and sequence repetition.
- *Forward Common Pilot Channel* is an unmodulated signal sent by AN to provide a phase reference for coherent modulation and a mean for signal strength comparisons between ANs for determining when to perform handover. No CRC and FEC is applied.
- *Forward Channel Quality Indicator Pilot Channel* is used by an AN to provide a reference for the measurement of the signals from the various antennas. No CRC and FEC is applied.
- *Forward Dedicated Pilot Channel* is an unmodulated signal transmitted by AN to provide coherent demodulation of BRCH channels. No CRC and FEC is applied.
- *Forward Acknowledgement Channel* is a portion of forward channel used for the transmission of acknowledgements from AN to multiple ATs for the data received on the Reverse CDMA/OFDMA Data Channel. No CRC and FEC is applied.

- *Forward Start of Packet Channel* is used on FL to indicate whether a persistent assignment is still valid or not. No CRC and FEC is applied.
- *Forward Shared Control Channel* is used on the FL to carry control information for the Forward Data Channel transmission. CRC-16 for non-GRA block and CRC-5 for GRA block are used with Rate 1/3 Convolutional encoding or Rate 1/3 Tail-biting convolutional encoding together with channel interleaver and sequence repetition.
- *Forward Pilot Quality Indicator Channel* is used to indicate the strength of the RL for a given AT. No CRC is used but Rate 1/3 concatenated encoding is applied.
- *Forward Fast Other Sector Interference Channel* is used on FL to carry indication of other sector interference. No CRC is used but Rate 1/3 concatenated encoding is applied.
- *Forward Reverse Activity Bit Channel* is used to carry a single bit to indicate the load on the RL for an AT. No CRC is used but Rate 1/3 concatenated encoding with sequence repetition is applied.
- *Forward Interference Over Thermal Channel* is used to indicate the interference levels in a given RL hop-pot subzone to AT in other sectors. No CRC is used but Rate 1/3 Concatenated encoding is applied.
- *Forward Power Control Channel* carries commands for closed-loop control of RL transmit power. No CRC and FEC is applied.
- *Forward Data Channel* carries higher level data and control information from AN to AT. CRC-24 is used with Rate 1/3 Convolutional encoding or Rate 1/5 Turbo encoding, or LDPC. Channel interleaver and sequence repetition (except for packet data control assignment block) are also applied.
- *Forward Broadcast and Multicast Services Channel* is used to carry broadcast and multicast data. CRC-24 and Rate 1/5 Turbo encoding is used with channel interleaver.
- *Forward Superposed Data Channel* is used to carry higher level data and control information that is sent as superposed traffic on the broadcast and multicast section. CRC-24 is applied with Rate 1/3 Convolutional encoding or Rate 1/5 Turbo encoding, or LDPC. Channel interleaver and sequence repetition are also applied.
- *Forward Broadcast and Multicast Pilot Channel* is an unmodulated signal transmitted by AN to provide phase reference for coherent demodulation of the Forward Broadcast and Multicast Channel. No CRC and FEC is applied.
- *Forward Superposed Channel Quality Indicator Pilot Channel* is used to feedback channel quality for the superposed traffic on broadcast and multicast section. No CRC and FEC is used.
- *Forward Superposed Dedicated Pilot Channel* is used for the channel estimation for the superposed traffic on the broadcast and multicast segment. No CRC and FEC is used.

Channels for RL are as follows:

- *Reverse Pilot Channel* is an unmodulated signal transmitted in the Reverse Link CDMA segment by AT to provide phase reference for coherent demodulation and signal strength measurements. No CRC and FEC is applied.

- *Reverse Auxiliary Pilot Channel* is an unmodulated signal transmitted in the Reverse Link CDMA segment in conjunction with Reverse CDMA Data Channel. It provides phase reference for coherent demodulation for demodulation of the Reverse CDMA Data Channel. No CRC and FEC is used.
- *Reverse Access Channel* is used to communicate to AN over a slotted random access scheme. No CRC and FEC is used.
- *Reverse CDMA Dedicated Control Channel* carries control information and other feedback information. No CRC and FEC is used.
- *Reverse CDMA Data Channel* is used to carry higher level data and control information from AT to AN. CRC-24 with Rate 1/5 Turbo or Rate 1/3 Convolutional encoding is used together with channel interleaver and sequence repetition.
- *Reverse Dedicated Pilot Channel* is an unmodulated signal transmitted in RL OFDMA segment by AT to provide phase reference for coherent demodulation of the Reverse OFDMA traffic channels. No CRC and FEC is used.
- *Reverse OFDMA Dedicated Control Channel* is a portion of a OFDMA Reverse Link to carry control information and other feedback information from AT to AN. CRC-9 and Rate 1/3 Convolutional encoding is used with channel interleaver and sequence repetition.
- *Reverse Acknowledgement Channel* is a portion of Reverse OFDMA channel to transmit acknowledgements from AT to multiple ANs. No CRC and FEC is used.
- *Reverse OFDMA Data Channel* is a portion of OFDMA Reverse Link to carry higher level data and control information from AT to AN. CRC-24 with Rate 1/5 Turbo or LDPC or Rate 1/3 Convolutional encoding is used together with channel interleaver and sequence repetition.

12.6.2 Coding and Modulation

Coding and modulation applies to both the FL and RL and is similar for both TDD and FDD modes. Subpackets are created from input packet if it is larger than the maximum subpacket size. Some channels have CRC encoding, and CRC-padded subpackets are encoded with one of the following:

- Rate 1/3 Convolutional encoding for block lengths $\leq$ 128bits
- Rate 1/5 Turbo encoding for block lengths $>$ 128 bits
- Rate 1/3 Tail-biting Convolutional encoding
- Rate 1/3 Concatenated encoding
- Optional Low Density Parity Check encoding for very high data rates

After encoding, channel interleaving and/or repetition is performed for some of the channels. Then, data scrambling is performed before modulation. The Forward Data Channel, the Forward Broadcast and Multicast Services Channel, the Forward Superposed Data Channel, and the Reverse OFDMA Data Channel shall use QPSK, 8-PSK, 16QAM, and 64QAM modulation. It also supports hierarchical modulation in which two modulation schemes are superimposed. For Broadcast and

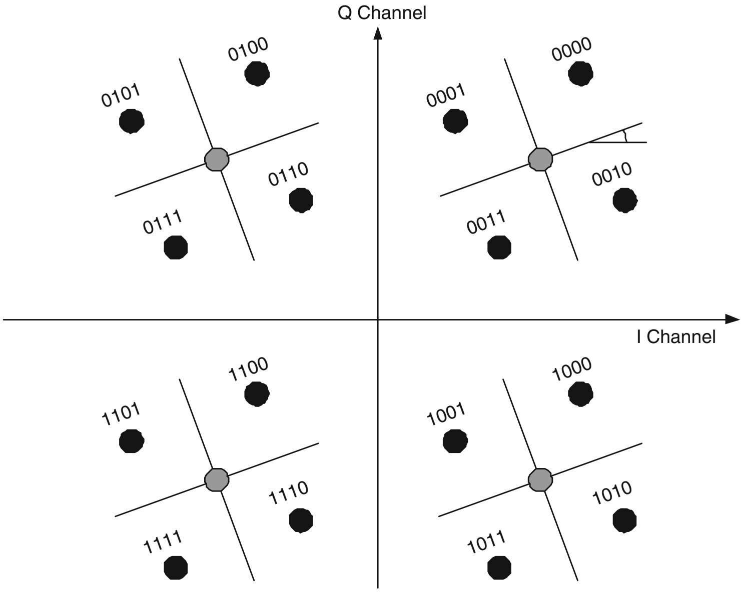


Fig. 12.7 Layered modulation: QPSK enhancement layer over QPSK base layer

Multicast Services, a QPSK enhancement layer may be superposed on a base QPSK or 16QAM layer to obtain the resultant signal constellation. The enhancement layer is rotated by the angle Θ in the counterclockwise direction as seen in Fig. 12.7.

12.6.3 OFDM Structure and Modulation Parameters

The frame structure in FL is divided into units of 25 superframes preceded by a superframe preamble as seen in Fig. 12.8. A superframe is made of OFDM symbols with a set of parameters as seen in Table 12.1. An OFDM symbol consists of N_{FFT} subcarriers, subcarriers; indexed 0 through $N_{\mathrm{GUARD,LEFT}} - 1$ and subcarriers indexed $N_{\mathrm{FFT}} - N_{\mathrm{GUARD,RIGHT}}$ through $N_{\mathrm{FFT}} - 1$ are not modulated.

The superframe begins with a superframe preamble, which consists of 8 OFDM symbols, indexed 0 through 7. The first OFDM symbol in the superframe preamble is transmitted over the Primary Broadcast Control Channel and next four OFDM symbols over the Secondary Broadcast Control Channel and the Quick Paging Channel in alternate superframes. The last three of these OFDM symbols are TDM pilots, indexed TDM Pilot 1, TDM Pilot 2, and TDM Pilot 3 as seen in Fig. 12.9.

UMB employs hierarchical search; TDM Pilot 1 is used for initial timing and frequency acquisition. TDM Pilot 1 forms the Forward Acquisition Channel and is unique across deployment by transmitting the same waveform by all eBSs. TDM

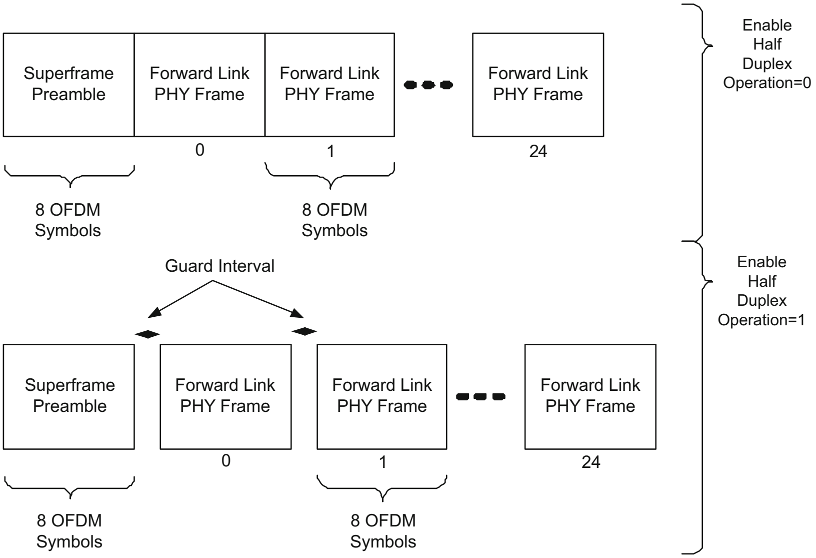
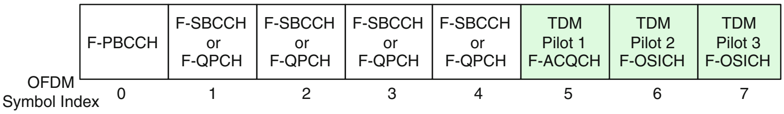


Fig. 12.8 Forward link superframe structure

Table 12.1 Forward and Reverse Link OFDM Symbol Numerology with respect to FFT size (N_{FFT})

Parameter	128	256	512	1,024	2,048
Chip rate (Mcps) $1/T_{\mathrm{CMIP}}$	1.2288	2.4576	4.1952	9.8304	19.6608
Subcarrier spacing (KHz) $1/(T_{\mathrm{CMIP}}N_{\mathrm{FFT}})$	9.6	9.6	9.6	9.6	9.6
Bandwidth of operation (MHz)	$B \leq 1.25$	$1.25 < \mathrm{B} \leq 2.5$	$2.5 < \mathrm{B} \leq 5$	$5 < \mathrm{B} \leq 10$	$10 < \mathrm{B} \leq 20$
Cyclic prefix duration (µs) T_{CP}= $(N_{\mathrm{CP}}N_{\mathrm{FFT}} \times T_{\mathrm{CHIP}}/16)$ $N_{\mathrm{CP}} = 1, 2, 3,$ or 4	6.51, 13.02, 19.53, or 26.04	6.51, 13.02, 19.53, or 26.04	6.51, 13.02, 19.53, or 26.04	6.51, 13.02, 19.53, or 26.04	6.51, 13.02, 19.53, or 26.04
Windowing guard interval (µs) $T_{\mathrm{WGI}} = N_{\mathrm{FFT}}T_{\mathrm{CHIP}}/32$	3.26	3.26	3.26	3.26	3.26
OFDM symbol duration (µs) T_{s}= $N_{\mathrm{FFT}}T_{\mathrm{CHIP}}(1 + N_{\mathrm{CP}}/16 + 1/32)$ $N_{\mathrm{CP}} = 1, 2, 3,$ or 4	113.93, 120.44, 126.95, or 133.46	113.93, 120.44, 126.95, or 133.46	113.93, 120.44, 126.95, or 133.46	113.93, 120.44, 126.95, or 133.46	113.93, 120.44, 126.95, or 133.46



OFDM Symbol Index	0	1	2	3	4	5	6	7
	F-PBCCH	F-SBCCH or F-QPCH	F-SBCCH or F-QPCH	F-SBCCH or F-QPCH	F-SBCCH or F-QPCH	TDM Pilot 1 F-ACQCH	TDM Pilot 2 F-OSICH	TDM Pilot 3 F-OSICH

Fig. 12.9 FL superframe preamble

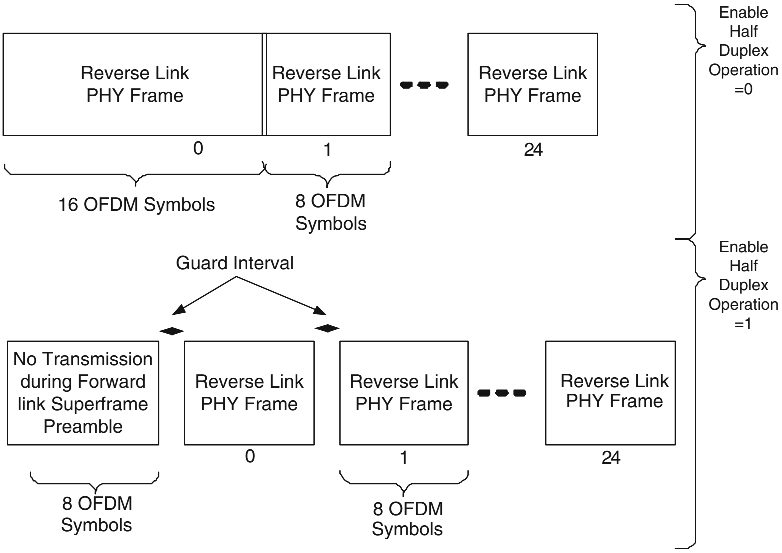


Fig. 12.10 Reverse link superframe structure

Pilot 1 uses a *Generalized Chirp-Like sequence*, which has low peak-to-average-power ratio. It is used to locate the superframe preambles transmitted by different eBSs. This search is windowed with one to two OFDM symbols in order to periodically perform the search with TDM Pilot 2 to find the other sectors. TDM Pilot 2 is used to search the other sectors regarding the TDM Pilot 1 search since TDM Pilot 2 carries PilotPN (Sector ID). This can be performed during initial network entry or handover. TDM Pilot 3 carries information such as synch/asynch, full/half duplex, etc. These are used to determine the system parameters. TDM Pilot 2 and TDM Pilot 3 are time domain sequences, which are chosen to be Walsh sequences with PN scrambling. There are 512 different Walsh sequences that allow 512 different Sector IDs. The corresponding frame structure for RL is seen in Fig. 12.10. Note that if half duplex operation is employed, FL superframe preamble time is silenced in RL.

Each symbol has a OFDM symbol start time determined by the superframe index with the superframe duration. Modulated complex symbols are converted to complex baseband waveform with inverse Fourier transformation. Then, windowing is performed before overlap-and-add operation, which adds OFDM symbols together to create the final complex baseband waveform.

12.6.4 HARQ

UMB supports synchronous HARQ on FL and RL. HARQ interlace structure may vary for different MAC packets. During the assignment of resources, interlace information is specified. Interlacing structures structure the timing relationship of retransmissions and acknowledgements. FL interlacing structures assigned via F-SCCH are as follows:

- *Eight interlace structure without extended transmission:* When a MAC packet arrives, first HARQ is transmitted in the corresponding frame, say k-th frame. The next one is then transmitted in $k+8n$-th frame and the ACK is received in $k+8n+5$-th frame when n stands for the transmission index.
- *Six interlace structure without extended transmissions:* Similarly, HARQ transmission is initiated following six PHY frames preceding HARQ transmission and the ACK is received three frames after the transmission.
- *Eight interlace structure with extended transmission:* Each HARQ transmission spans three PHY frames as seen in Fig. 12.11. For example, first HARQ transmission may be initiated at k-th, $k+1$-th, $k+2$-th frames and ACK is sent in $k+5$-th frame. The next ones are then sent in the $k+8n$-th, $k+8n+1$-th, $k+8n+2$-th frames and ACK is sent in $k+8n+5$-th frame.
- *Eight interlace structure with two frame transmission:* It is mandatory in AT when operation is full-duplex. Each HARQ transmission of a data packet spans up to two consecutive PHY frames in this case. For instance, first HARQ transmission may use frame k-th or $k+1$-th or both. ACK is received in $k+5$-th or $k+6$-th. Next HARQs are then sent in frames at $k+8n$-th and $k+8n+1$-th frames.
- *Six interlace structure with two frame transmission:* It is optional in the AT. This is similar to eight interlace structure with two frame transmission but transmissions are interlaced with six frames.

RL interlacing structure is similar and only supports eight interlace structure with and without extended transmissions. The size of a MAC packet is a function

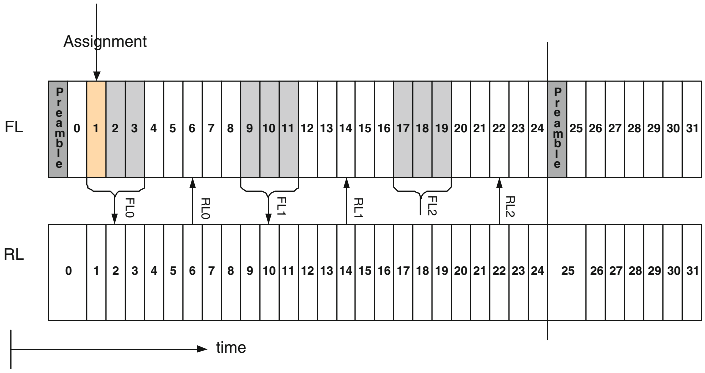


Fig. 12.11 FL eight interlace structure with extended transmissions

of the packet format (PF) and the number of subcarriers that are assigned to the data packet. Packet format specifies the spectral efficiency used on the first HARQ transmission and the modulation format to be used for each HARQ transmission. A packet format (PF) specifies the modulation orders of 2, 3, 4, and 6, which correspond to QPSK, 8PSK, 16QAM, and 64QAM.

12.6.5 Multiple Antenna Procedures

UMB simultaneously supports multiple antenna operations including MIMO, SDMA, and beamforming. MIMO procedures consist of (a) single-code word (SCW) or multicode word (MCW) operations, (b) precoding, (c) permutation matrices for multicode word MIMO and/or SDMA. There are multiple *physical antennas* and a *logical antenna* is defined as a linear combination of physical antennas that is slowly varying over time and frequency. Also, some of the logical antennas are indexed as an *effective antenna* and each effective antenna is associated with Forward Common Pilot Channel and Forward Channel Quality Indicator Pilot Channel. Transmission in Forward Data Channel is performed on effective antennas, or linear combinations of effective antennas. Also, fixed linear combination of physical antennas constitutes a tile antenna. Combination may vary arbitrarily from tile to tile. The Forward Data Channel may use tile-antenna if the resource is being modulated in block zone and use effective antenna without precoding or logical antenna with precoding in distributed zone according to the resource multiplexing mode.

Two types of precoding codebooks are supported: Knockdown Codebook and Readymade Codebook. The mapping between effective antennas and logical antennas used for precoding depends on the type of the codebook:

- *Knockdown Codebook* is constructed with vectors from a selected universal matrix. The AT selects one of the two universal unitary matrices according to the preferred matrix index. From the preferred matrix, the AT selects the preferred column, which is indicated by the vector bitmap. There are two kinds of default knockdown codebooks: binary unitary codebook, which is defined by the 4×4 identity matrix I_4 and fourier matrix-based codebook, which is defined by $M \times M$ matrices $(H_M^{(g)})$; $H_M^{(g)}$ are defined as follows:

$$H_M^{(g)} = [h_{nm}^g] = [\mathrm{e}^{j\frac{2\pi n}{M}(m+\frac{g}{G})}], \tag{12.1}$$

  where $m,n = 0,1,\ldots,M-1$.
- *Readymade Codebook* defines 64 precoding matrices. The AT selects one of the matrices according to the precoder index. The Precoder is constructed by the first r vectors where r is the required rank, indicated by Channel Quality Indicator for MCW MIMO or by the rank feedback in an explicit manner for SCW MIMO.

In addition to default knockdown codebooks, codebooks can be configurable with download mechanism. Also, precoding matrices in a codebook may be grouped

Table 12.2 MIMO performance analysis

$Tx \times Rx$	FL (Mbps)	RL (Mbps)
1×2	11.9	8.1
1×4	15.2	13.9
2×2	12.7	
4×2	13.2	
4×4	21.0	

Sector capacity with 10MHz, FDD mixed channel model, 16 users/ sector, full buffer traffic, proportional fair scheduling, 2.0 km site-to-site distance, SCW MIMO precoding in FL, results given after subtracting all overhead. The results do not include subband scheduling gains. VoIP users per sector is found to be more than 500 with Enhanced Variable Rate CODEC and dual Rx eBS (source: www.cdg.org)

into clusters where a cluster spans only part of the space and the other cluster can be used to form spatial beams as in SDMA, which allows transmitting different symbols to different ATs through the same subcarrier.

MIMO gain is listed in Table 12.2, which depends on network load. In general, gain decreases as the load increases depending on the channel. A 2 × 2 MIMO scheme gives 10–50% capacity gains and a 4 × 4 MIMO scheme gives 60–120% capacity gains.

12.6.6 Hop-Port Definition and Indexing

Now, we start introducing the basic foundation for resource assignment. A hop-port is defined as a resource where hop-port indexing maps the subcarriers to hop-ports. Each OFDM symbol consists of $Q_{\mathrm{SDMA}}N_{\mathrm{FFT}}$ individually modulated hop-ports where Q_{SDMA} is equal to SDMA dimensions, if any. The hop-permutation maps the $Q_{\mathrm{SDMA}}N_{\mathrm{FFT}}$ hop-ports to N_{FFT} subcarriers for each Q_{SDMA} *subtrees*.

In RL, hop pattern generation is a two-step process: mapping hop-ports to subcarriers regarding the nominal location of CDMA subsegments, which contains 128 contiguous subcarriers, and relocating the subcarriers when CDMA subsegments hop among the CDMA hopping zones. The Reverse OFDMA Data Channel supports global hopping (GH) and local hopping (LH). The primary difference between GH and LH structures is that in the LH structure, a hop-port hops within a subzone as seen in Fig. 12.12, while in the GH, a hop-port may hop over the entire bandwidth as seen in Fig. 12.13.

In FL, Forward Data Channel supports Distributed Resource Channel (DRCH) and Block Resource Channel (BRCH). With DRCH, a set of hop-ports is mapped to subcarriers that are scattered across the entire bandwidth or across a large subset of the bandwidth. Subcarriers are regularly spaced. BRCH structure assigns a

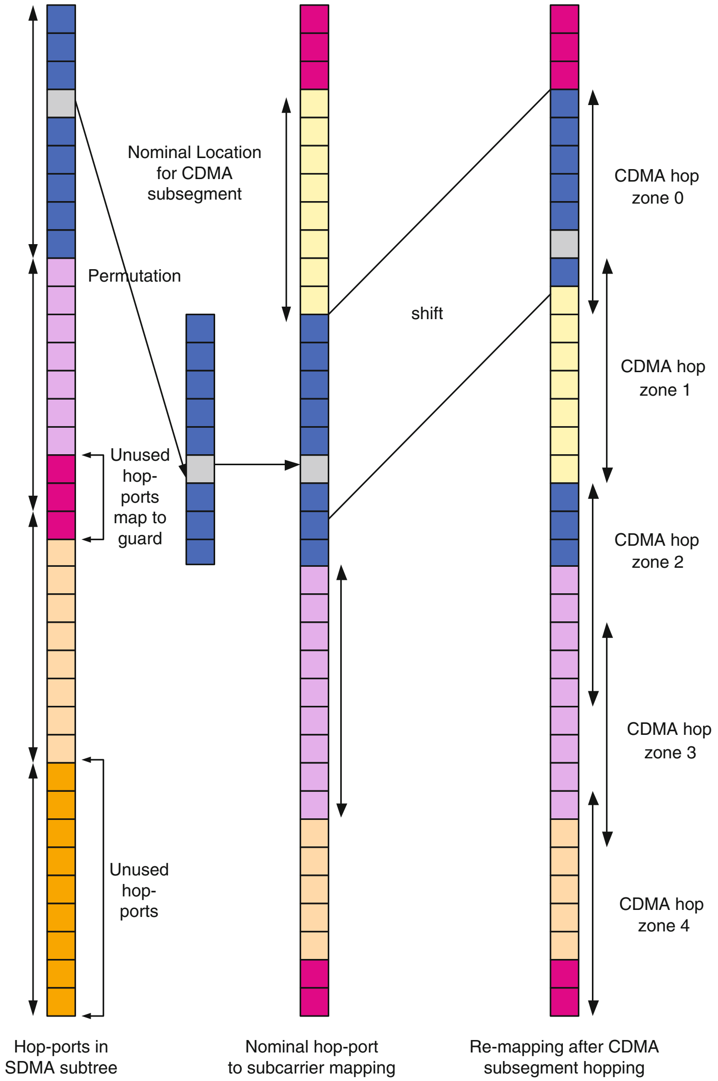


Fig. 12.12 Hop-port to subcarrier mapping for the LH

set of contiguous subcarriers, and mapping between hop-ports and frequency is kept constant through the PHY frame as seen in Fig. 12.14. To leverage both frequency diversity and frequency-selective (multi-user diversity) transmission, both modes can be combined within a PHY Frame as seen in Fig. 12.15 where there are two multiplexing modes[3]; first mode punctures DRCH onto BRCH structures and second mode uses them in different zones.

[3] ResourceChannelMuxMode.

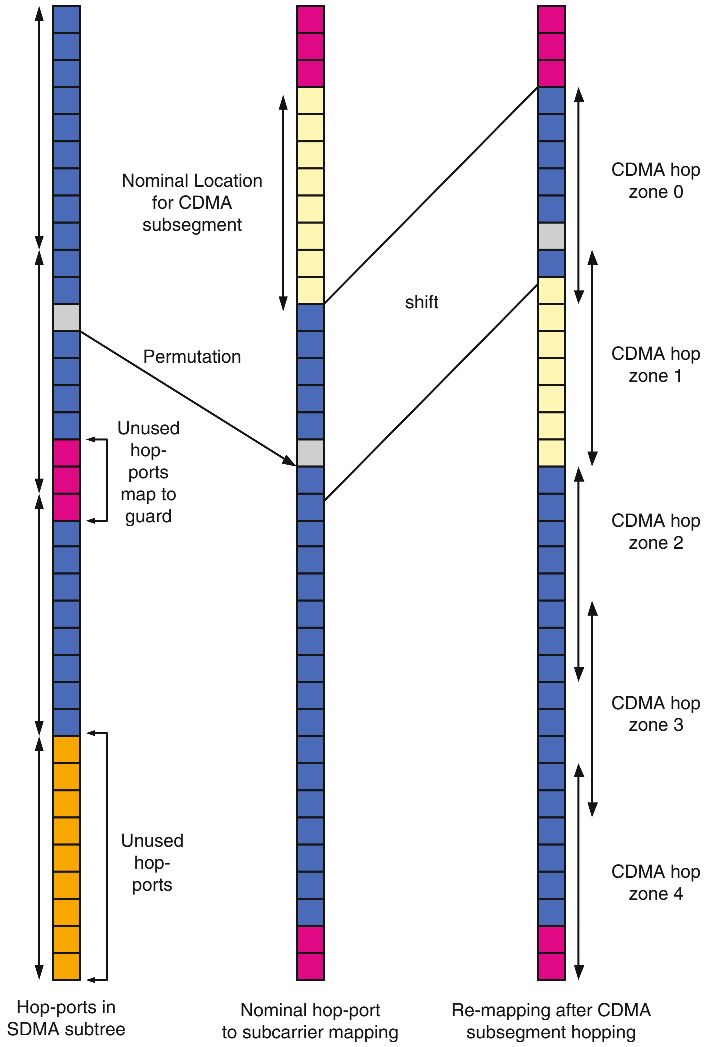


Fig. 12.13 Hop-port to subcarrier mapping for the GH

12.6.7 Channel Tree

A channel tree associates sets of hop-ports to node identification number (NodeID) where hop-port is the fundamental unit of resource assignment as seen in Fig. 12.16. Each hop-port maps to one unique subcarrier. Mapping of hop-ports to subcarriers is specified by the physical layer.

Nodes define orthogonal assignments; if a node is in use, then all descendants and ancestors of the node become unavailable for use and so are called “restricted” nodes. In case of SDMA transmission, where a subcarrier is used at different spatial

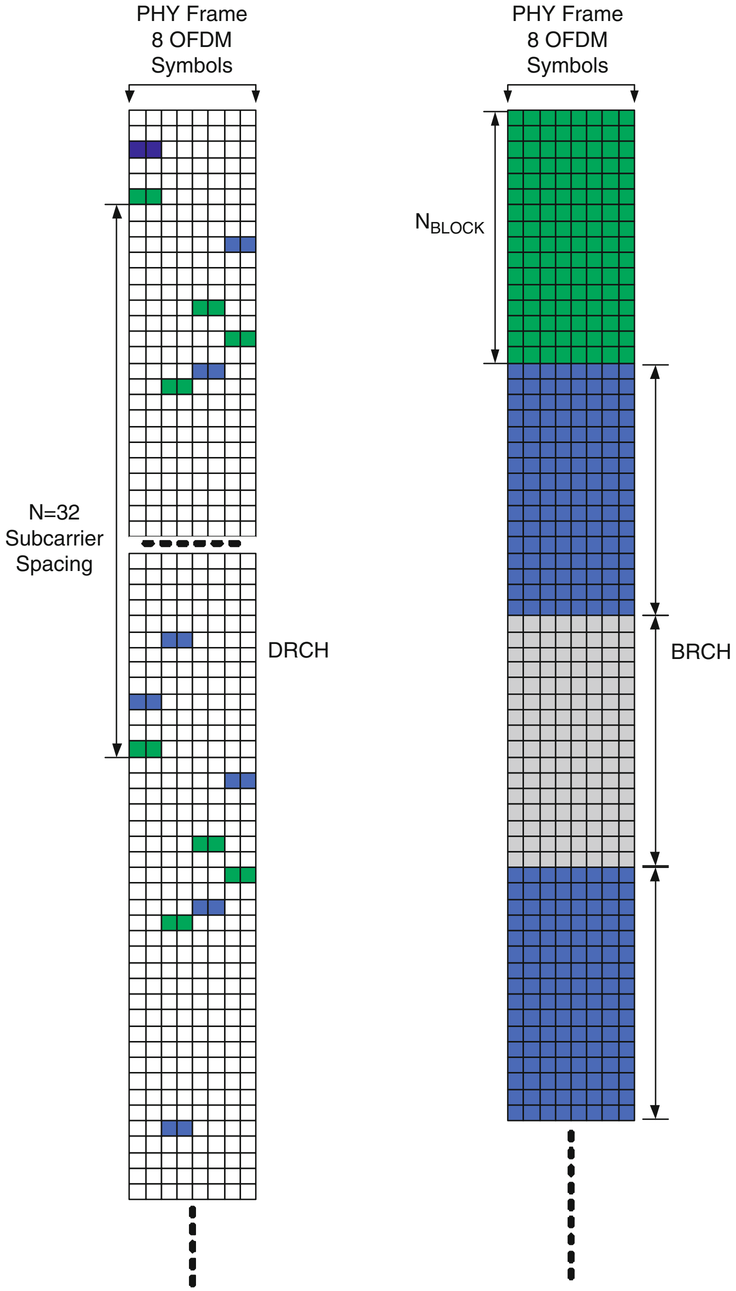


Fig. 12.14 Forward data channel resource allocation: distributed resource channel and block resource channel (© 3GPP2)

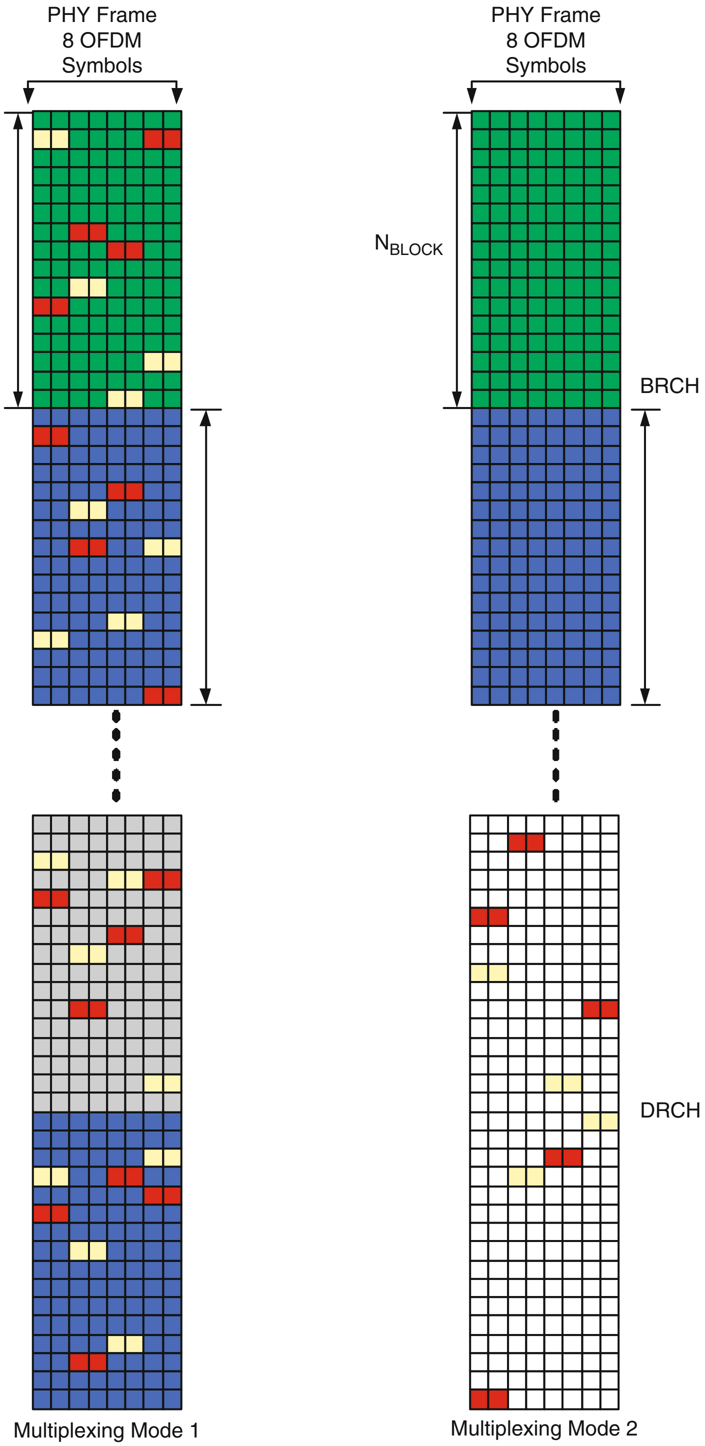


Fig. 12.15 The forward data channel resource allocation: multiplexing mode (© 3GPP2)

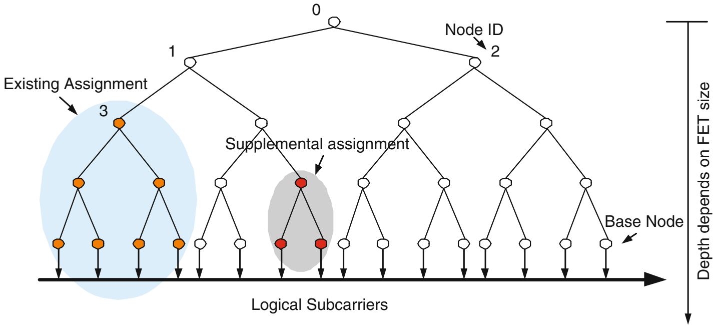


Fig. 12.16 Channel tree

dimensions, a SDMA subtree is defined to specify all associated hop-ports. Hop-ports from different subtrees may map to the same subcarrier.

12.6.8 Resource Management

Resource allocation is centralized at eBS for both FL and RL. Scheduler in eBS determines the FL rate based on FL channel quality reports from terminal and it considers channel feedback as well as resource requests for the RL rate determination. Scheduler implements algorithms to

- Maximize system capacity
- Manage QoS requirements such as mobile throughput and latency
- Maintain fairness across mobiles with widely disparate channel qualities
- Ensure that the scheduler has information required to utilize features such as subband scheduling, fractional frequency reuse, precoding, and SDMA to achieve the aforementioned goals

The minimum unit of resource is called *tile*, which is a group of 16 hop-ports for the duration of eight OFDM symbols. Tiles for FL and RL are assigned to a user by eBS using F-SCCH. F-SCCH may have more than one Link Assignment Block (LAB). LAB indicates the MAC ID of the destination AT and can be of various types; examples are FLAB for FL assignment, RLAB for RL assignment, SCW-FLAB, MCW-FLAB, etc.

LAB blocks are distinguished by block headers and may include the following:

- Node ID in the channel tree for the tiles as seen in Fig. 12.16
- Packet format on the assigned tiles
- Tx power assigned to AT
- Pilot format
- HARQ interlace structures
- MIMO rank and precoding information

- Persistent or nonpersistent assignment
- Supplemental assignments that can be used to add or remove tiles, and also change the PF, pilot format, rank, etc. of an existing assignment

These assignments can be of two types: *persistent* or *non-persistent*. Persistent assignments reduce the overhead of bandwidth request by allocating continuously LABs until it is explicitly deassigned. Nonpersistent assignment on the other hand maximizes the utilization adaptively for delay insensitive traffic. Assignment becomes obsolete when the packet is transmitted successfully or after six HARQ transmissions.

Resource assignment can be modified during the HARQ: the Node ID of an assignment can be changed in the middle of HARQ retransmissions of a packet to defragment the channel tree resources. However, other fields cannot be changed. Also, number of HARQ transmissions can be increased with HARQ extension. Each extension gives six more HARQ retransmissions and can be used multiple times.

UMB also offers residual-resource assignment (RRA). RRA assigns unused frames of a persistent allocation of AT to a different AT. RRA is useful for VoIP applications where there is silence suppression. They are nonpersistent allocations and expire right after packet decoding.

12.6.9 Interference Management

As we discussed earlier, UMB offers optional support for disjoint links in order to have AT independently select the strongest FL and strongest RL sectors, which may reside in different eBSs. eBSs monitor the pilot strength in the RL and reports to the AT. AT is power controlled by the strongest RL sector. This helps to alleviate the interference in the cell edges along with *fractional frequency reuse*. Fractional frequency reuse utilizes the entire bandwidth in the inner cell and fraction of the bandwidth at the cell edge. We presented this type of allocation in WiMAX as well. This provides flexibility in terms of distinguishing users according to their signal characteristic. Scheduler dedicated to FL assigns users closer to the eBS to subcarriers with low power, and cell-edge users to subcarrier with high power. This increases sector capacity without changing the cell-edge data rates.

RL interference management in UMB considers two prong ways: regular and fast interference management. Regular interference management considers other cell interference to enable universal frequency reuse. In loaded scenarios, transmit power of different ATs is structured in a way that the amount of interference they cause to neighbor cells is minimized. This interference control is tightly applied to keep the RL stable. In regular RL interference management, eBSs broadcast an interference status indication via F-OSICH, which is high/low indication. eBSs that are active set members of an AT feed back RL pilot strength (R-PICH) using F-PQICH. AT then computes the relative strength of each eBS with respect to RL of serving eBS. Every AT runs a closed loop based on *ChanDiff* and F-OSICH where smaller ChanDiff and larger transmit power reduce their power aggressively. The value of this transmit power is sent to the serving eBS in order to assign this power level to AT with a suitable rate regarding buffer size, power headroom, QoS level, etc., as well.

Fast RL interference management considers dynamic control of power levels in bursty traffic situations. Every eBS also broadcasts a busy bit via F-FOSICH based on interference observed on every subzone of every RL frame. This indicates the load in a given subzone of a RL frame in other sectors. When AT receives a busy bit for the same subzone it is transmitting, it brings down its transmit power. AT also reports this change to eBS and eBS adjusts the power of others since it knows that AT is affected by that F-FOSICH. eBS also broadcasts interference level seen in every subzone so that AT uses it in its power calculations to reduce packet errors.

12.6.10 Power Savings

Power saving option in UMB comprises quick paging, selected-interlace state, and semiconnected state:

- *Quick Paging* reduces the page period and the energy spent during each page. These two determine the power consumption of AT in idle state. For every 50ms, a quick page block is transmitted. AT needs to be awake ∼0.5ms to decode five OFDM symbols to get a quick page. Less frequent paging improves the standby time, and frequent paging reduces the latency during call setup.
- *Selected-interlace state* implements a power saving mechanism, which helps an AT to turn off its receiver on chosen interlaces. Interlaces are negotiated beforehand and assignments are done only on selected interlaces. RL control channels are active. AT wakes up to receive power control commands and measure CQI pilot. AT is also allowed to send RL control and RL data when it wakes up. eBS can further optimize power savings by selecting the interlace on/near the RL control segment interlace, slowing down the CQI/pilot reporting period, and matching power control commands sent with selected interlaces. This power saving state can be useful with low rate traffic such as VoIP.
- Semiconnected state keeps FL active but avoids RL control transmission. It is negotiated and AT keeps its MAC ID in this state and can easily shift to fully-connected state with minimal delay (less than 30ms) via F-SCCH sent from eBS. It does not require paging. This method may result in huge savings during bursty traffic such as web browsing.

12.7 Summary

UMB integrates advanced radio access techniques into a single global standard in order to fulfill 4G requirements for 3GPP2. UMB is based on a flat All-IP network architecture as in WiMAX and LTE, called Converged Access Network (CAN). The CAN architecture simplifies the core network and offers PPP-free operation with simpler interfaces and enables load balancing across the network elements. CAN provides seamless handover to EV-DO and 1x networks and other non-3GPP2 technologies. Network elements include Access Gateway (AGW), Session Reference Network Controller (SRNC), evolved Base Station (eBS), and Access Terminal (AT). AGW is responsible of maintaining the data path functionality. It is the anchor mobility point for mobility within AGW domain and assists Home Agent (HA) for mobility across AGW domains. SRNC is responsible for facilitating control signaling. eBS implements UMB air interface like AT and maintains the over-the-air communication with AT. eBS also performs QoS classification, admission control, and scheduling.

Multiroute functionality is a distinguished feature to enable fast switching between eBSs with low overhead. Multiroute enables AT to establish routes with multiple eBSs at the same time and AT chooses eBSs for its forward and reverse links independently. AT also has the flexibility to change the forward and reverse links independently for layer 1/2 handover. Layer 3 handover is also disassociated from layer 1/2 handover to enable fast and efficient mobility across eBSs.

UMB air interface offers flexible and scalable bandwidth support from 1.25 to 20 MHz in steps of $\sim$154 KHz. FDD/TDD modes are supported and TDD standardization is in progress. UMB FDD is proposed as FDD mode of IEEE 802.20 (Mobile Wireless Broadband Access) standard (See last chapter). It is designed for full frequency reuse that does not need frequency planning. Resource allocation offers low latency and introduces persistent and nonpersistent allocation. The UMB control mechanisms optimize the transmission of variable length packets for each application based on the QoS requirements of each application and user.

References

1. “Overview for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-000-0 v2.0, 2007. http://www.3gpp2.org.
2. “Physical Layer for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-001-0 v2.0, 2007. http://www.3gpp2.org.
3. “Medium Access Control Layer for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-002-0 v2.0, 2007. http://www.3gpp2.org.
4. “Radio Link Layer for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-003-0 v2.0, 2007. http://www.3gpp2.org.
5. “Application Layer for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-004-0 v2.0, 2007. http://www.3gpp2.org.
6. “Security Functions for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-005-0 v2.0, 2007. http://www.3gpp2.org.
7. “Connection Control Plane for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-006-0 v2.0, 2007. http://www.3gpp2.org.
8. “Session Control Plane for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-007-0 v2.0, 2007. http://www.3gpp2.org.
9. “Route Control Plane for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-008-0 v2.0, 2007. http://www.3gpp2.org.
10. “Broadcast-Multicast Upper Layer for Ultra Mobile Broadband (UMB) Air Interface Specification,” TSG-C, C.S0084-009-0 v2.0, 2007. http://www.3gpp2.org.

11. “CAN Wireless IP Network Overview and List of Parts,” TSG-X, X.S0054-000-0 v1.0, 2007. http://www.3gpp2.org.
12. “Basic IP Service for Converged Access Network Specification,” TSG-X, X.S0054-100-0 v1.0, 2007. http://www.3gpp2.org.
13. “Multiple-Authentication and Legacy Authentication Support for Converged Access Network,” TSG-X, X.S0054-102-0 v1.0, 2007. http://www.3gpp2.org.
14. “MIPv4 Specification in Converged Access Network Specification,” TSG-X, X.S0054-110-0 v1.0, 2007. http://www.3gpp2.org.
15. “CMIP Based Inter-AGW Handoff,” TSG-X, X.S0054-210-0 v1.0, 2007. http://www.3gpp2.org.
16. “Network PMIP Support,” TSG-X, X.S0054-220-0 v1.0, 2007. http://www.3gpp2.org.
17. “QoS Support for Converged Access Network Specification,” TSG-X, X.S0054-300-0 v1.0, 2007. http://www.3gpp2.org.
18. “Converged Access Network Accounting Specification,” TSG-X, X.S0054-400-0 v1.0, 2007. http://www.3gpp2.org.
19. “CAN Data Dictionary,” TSG-X, X.S0054-910-0 v1.0, 2007. http://www.3gpp2.org.

Chapter 13
Drivers of Convergence

The cellular access networking has undergone a major transition in the past decade with the convergence toward OFDMA and IP technologies. Access network technologies based on these introduced by WiMAX, 3GPP, and 3GPP2 are mastered in earlier chapters with in-depth understanding of those technologies, terminology, and network configurations. Interworking of these access networks with wireline technologies is a significant step to achieve a single telecommunications network foundation. Fixed-mobile convergence (FMC) addresses this network convergence together with service convergence and device convergence in order to provide convenience and simplicity for consumers and business users to get highly featured but lower cost communications.

Device convergence in FMC framework introduces diverse functionality in a single device and works together with network convergence to provide connectivity to services using the most suitable access technology at any location or moment in time. Finally, service convergence enables the delivery of services seamlessly and transparently to the user over any network.

Convergence will obliterate the physical barriers that now limit service providers from reaching all their customers with all types of services. It will let wireline service providers have the flexibility to leverage the wireless networks, while wireless network operators will use the most robust network resources available to accommodate growing demand from mobile subscribers. They will both have the ability to share resources and interact with each other in order to create new efficiencies.

We dedicate this last chapter to discuss briefly this fixed-mobile convergence trend in all three angles and talk about the future OFDMA and IP technologies that will contribute to the widespread availability of broadband access and the phenomenal growth of wireless networking. We first discuss interworking between access technologies covering WiMAX interworking to LTE, 3GPP, 3GPP2, and DSL as well as LTE interworking to HRPD (High Rate Packet Data Service) of 3GPP2 and WLAN in the context of Generic Access Network (GAN) framework. We also mention Media Independent Framework being designed in IEEE 802.21 standard. These interworking solutions are ongoing work and subject to change. Second, we follow this FMC discussion by describing the service and device convergence. We cover

M. Ergen, *Mobile Broadband: Including WiMAX and LTE*,
DOI: 10.1007/978-0-387-68192-4_13, © Springer Science+Business Media LLC 2009

Policy and Charging Control (PCC) and IP Multimedia Subsystem (IMS) in service convergence along with Open Mobile Alliance. Then we discuss Open Handset Alliance (OHA, www.openhandsetalliance.com), Open Base Station Architecture Initiative (OBSAI, www.obsai.org), and Advanced Telecommunications Computing Architecture (ATCA) as an examples to device convergence. In the second half of this chapter, we focus on new OFDMA-based technologies such as the following:

- IEEE 802.16j framework for Mobile Multihop Relay
- IEEE 802.16m framework for Advanced Air Interface to WiMAX
- IEEE 802.20 framework for Mobile Broadband Wireless Access
- IEEE 802.22 framework for Cognitive Radio

13.1 Network Convergence

Network convergence ultimately aims a single access network to accommodate multiple services. A great step toward that is interworking of current access technologies as seen in Fig. 13.1. Mobility among 3GPP, 3GPP2, and WiMAX is described in this section along with expanded reach through WLAN/WiFi and DSL. We focus more on WiMAX integration in which WiMAX is considered as a visited system for dual-mode devices homed in 3GPP/2 system. The network access and session mobility scenarios may range from simple integration to a sophisticated integration in order to enable session continuity as follows:

- Scenario 1: *Common Billing and Customer Care* would be the simplest scenario to be enabled if both technologies are owned by the same operator. This would have no impact on the network architecture.
- Scenario 2: *Access to the Internet* would require device/user authentication in the WiMAX system using its credentials. The realm portion of NAI will let AAA messages to be routed to 3GPP/2 home system. Access to the Internet would be through CSN of WiMAX network.
- Scenario 3: *Access to the Home 3GPP/2 Services* would require using the common Home Agent of the 3GPP/2 home system. MS may access any packet-based services through the HA (PDN GW in 3GPP).
- Scenario 4: *Session Continuity* requires maintaining the IP address when handing off from 3GPP/2 to a WiMAX interface. Seamless session continuity in dual-mode devices can be obtained by keeping both interfaces active for a period of time.
- Scenario 5: *Access to the Circuit Switched (CS) Services* may require listening the voice and other circuit switched services (SMS, location requests, etc.) even when it is in WiMAX network coverage. VoIP service in WiMAX can cannibalize CS voice services.

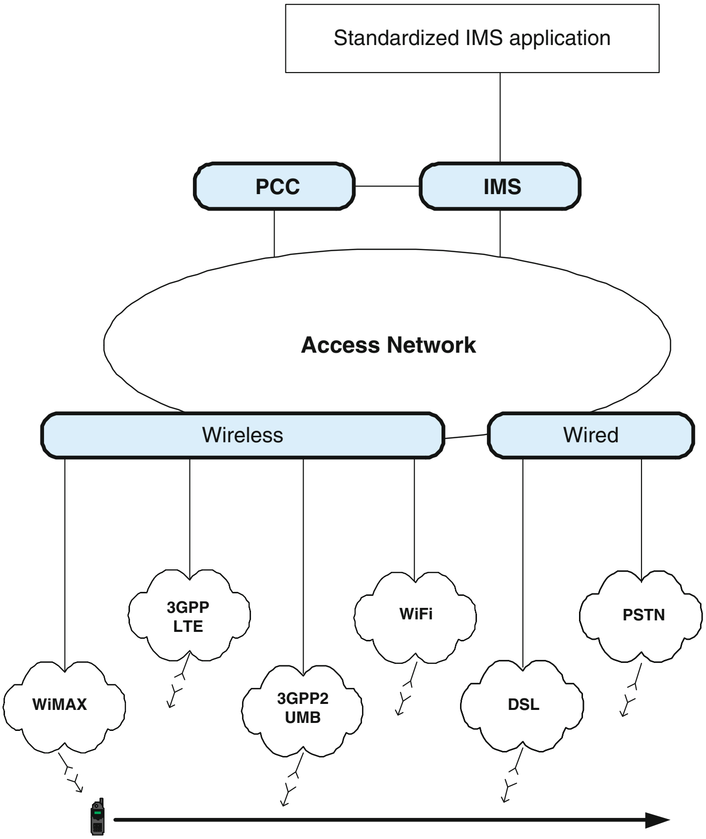


Fig. 13.1 Convergence networks

13.1.1 LTE Interworking with WiMAX

LTE framework in 3GPP is conducting a feasibility study to provide an interworking solution to WiMAX for scenario 4 [TR 36.938]. In scenario 4, mobility between E-UTRAN architecture of LTE and WiMAX should support bidirectional service continuity and seamless voice continuity. Figure 13.2 shows the reference architecture including Forward Attachment Function (FAF) and a new reference point X200 in the EPC architecture. UE uses X200 reference point to communicate to FAF over an IP access in order to establish preregistration or resource request preparation in the target access network. UE (MS) may also be entitled to receive access network information of supported WiMAX access technologies. E-UTRAN may transmit following key information for WiMAX neighbor cells:

- DL center carrier frequency (multiple of 250 KHz)
- Cell bandwidth
- Preamble index

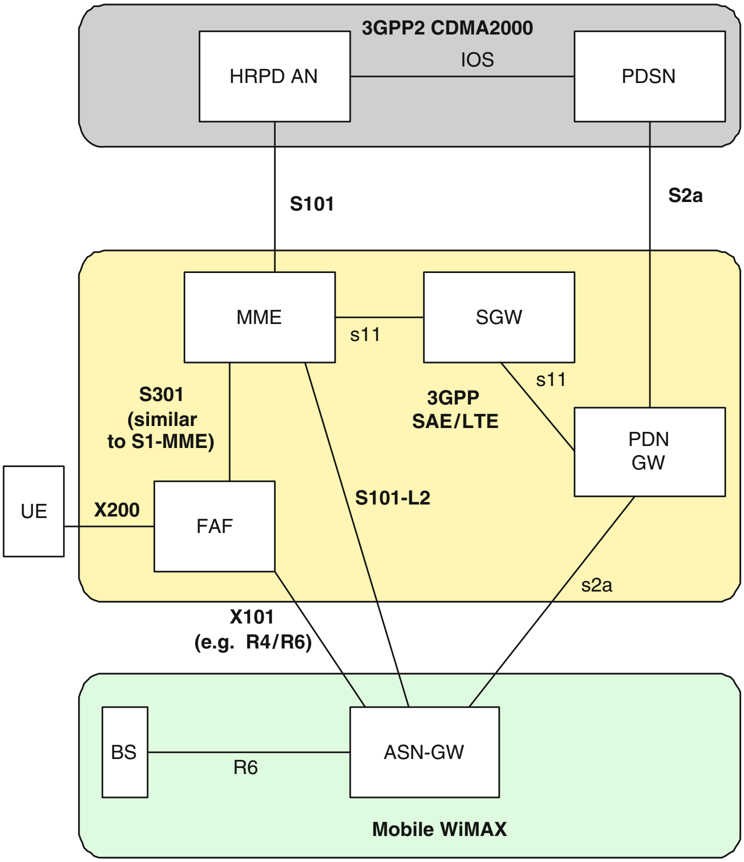


Fig. 13.2 LTE interworking with WiMAX and HRPD: X101 reference point may be based on R4 or R6 of WiMAX reference points

- BS ID/NAP ID/NSP ID
- MAC and System version
- Available DL and UL radio resources
- Cell size

UE may do measurement in idle or active mode for the target cells. This measurement report may contain RSSI and CINR values of neighboring cells and is reported to perform cell reselection. Preregistration to WiMAX system in advance for a handover reduces the time involved in the process. Preregistration is initiated by a signal from 3GPP network, which lets UE register to the target WiMAX network via 3GPP network. UE passes any essential context to the target network via L2 or L3 tunneling. In case of L2 tunneling, the source eNB forwards these messages to the MME, which sends them to the WiMAX ASN-GW as IP payload via bidirectional L2 tunnel. Handover to WiMAX is initiated with a decision to UE from 3GPP network.

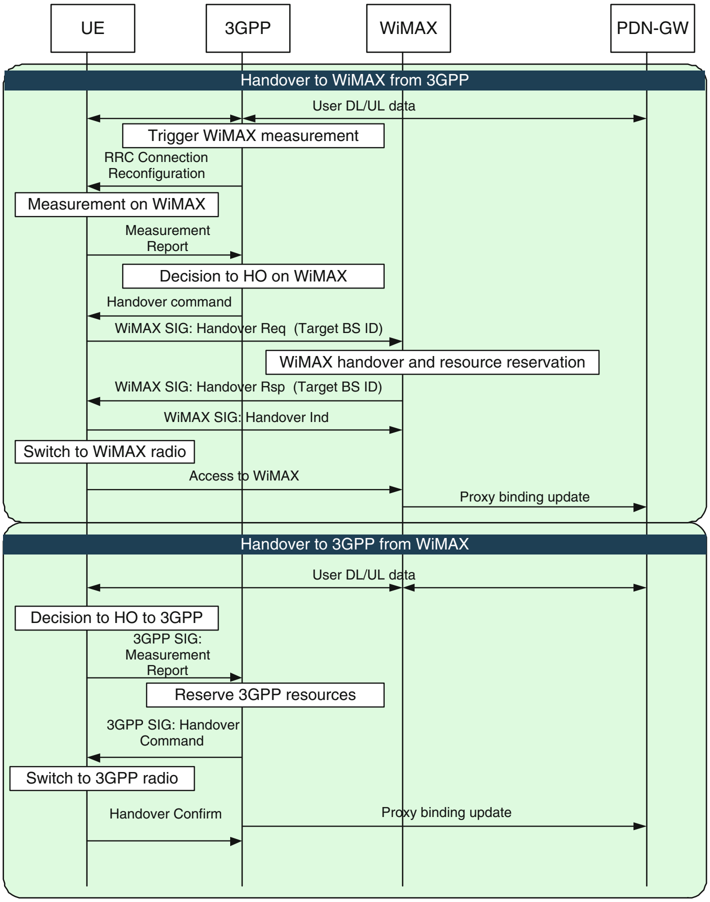


Fig. 13.3 Intertechnology handover between LTE and WiMAX

Handover request is sent from UE to WiMAX via 3GPP and WiMAX tunneling mechanism. WiMAX network establishes the context and responds with handover response to UE, which replies back with handover indication message. After handover indication, UE switches to the WiMAX radio. WiMAX network establishes binding update with PDN GW to update the anchor data path. The handover flow chart is depicted in Fig. 13.3.

13.1.2 LTE Interworking with HRPD of 3GPP2

Considering the existing CDMA2000 networks, LTE is being designed to provide interworking solution to provide seamless service continuity between CDMA2000

HRPD (1xEX-DO) Release 0 and A and E-UTRAN.[1] Figure 13.2 depicts the corresponding interfaces to HRPD network. CDMA2000 network supports handover to a E-UTRAN architecture by using the eNB and MME as relay points in order to relay CDMA2000 messages. The UE monitors the E-UTRAN BCCH or any other dedicated channel to retrieve the 3GPP2 system information.[2] The following information shall be transmitted on E-UTRAN:

- HRPD preregistration allowed
- HRPD preregistration zone
- HRPD neighbor bandclass
- HRPD neighbor frequency
- HRPD neighbor PN sequence offset
- HRPD pilot PN sequence offset index increment
- HRPD timing reference
- HRPD searching window size
- Number of HRPD neighbor bandclasses
- Number of HRPD neighbor PN sequence offsets
- HRPD start measuring E-UTRAN signal quality threshold
- HRPD start measuring E-UTRAN Rx power strength threshold

The measurement may be triggered in idle or active mode according to certain thresholds. UE may perform preregistration to HRPD system in advance to reduce the handover time. If preregistration zone changes, the UE may perform preregistration again. This procedure is transparent to E-UTRAN network. HRPD preregistration is performed through eNB and MME by using the generic messages. S1-AP messages are used to transfer these messages from eNB to MME and tunneled to HRPD access via a tunnel between MME and HRPD access. These tunnels are also utilized during handover. After HRPD handover decision, UE initiates the HRPD handover signaling to HRPD network via E-UTRAN entities. Handover is complete by sending a binding update to PDN GW for the new anchor data point. The data are then forwarded from PDN GW to HRPD access. During handover from HRPD to E-UTRAN, UE shall perform an "attach" procedure and trigger relocation to E-UTRAN via HRPD–MME interface while it is connected to HRPD. After handover signaling, the data path in the PDN GW is updated toward new SGW.

13.1.3 WiMAX Internetworking with 3GPP2

Currently, WiMAX and 3GPP2 interworking solution is motivated to provide an interworking solution for an operator that owns and operates both networks with dual-mode devices. The loosely coupled solution complies with the 3GPP2 S.R0087-A

[1] The 3GPP2 also introduces three specs under TSG-S: S.R0128 designs interworking between HRPD and WiMAX, S.R0129 defines HRPD and LTE, and S.R0130 defines UMB and LTE.

[2] "The HRPD system information block (SIB) shall be included in the Dynamic System Information (SI) of E-UTRAN BCCH." Source: 3GPP TR 36.938.

CDMA2000-WLAN Interworking and uses common core elements such as AAA, HA, DHCP servers, and PCC. The Client MIP is assumed to be common to WiMAX and CDMA2000 interfaces. The CMIP client may receive an agent advertisement from the FA of PDSN or ASN. The network selection may be automatic or manual in which user selection may override any selection criteria. Mobility keys are retrieved after device/user authentication and authorization in addition to the outer NAI. MIP client uses the pseudoIdentity@realm and HoA values for 3GPP2 and WiMAX MIP registration in order to maintain the original binding. HoA is notified back to MS with MIP registration response if dynamic address allocation is requested.

Also, mixed mode is provisioned as follows where WiMAX PMIP mode is interworked with 3GPP2 CMIP mode. If handover is from 3GPP2 CMIP to WiMAX PMIP mode, during device/user authentication to WiMAX network with NAI, H-AAA assigns the HA via RADIUS Access-Accept message, which is stored at home AAA during 3GPP2 CMIP registration. MS transmits DHCP Discover to ASN-GW, which initiates PMIP procedure from FA to the HA. The HA replies to FA with HoA in registration response message. The HoA is then sent to MS with DHCP OFFER message.

If handover is from WiMAX PMIP to 3GPP2 CMIP, then from the new CoA, MS CMIP sends registration request with NAI, which is relayed up to H-AAA via FA in PDSN. H-AAA returns with HA address, which is stored during WiMAX PMIP registration. The FA in PDSN then initiates MIP registration request to HA in order to update the binding of HoA with the new CoA. HoA, HA, and NAI are then relayed to CMIP in MS via FA.

13.1.4 WiMAX Internetworking with 3GPP

Previously, we mentioned about a more advanced and integrated interworking solution between WiMAX and 3GPP LTE, which is under development as part of the 3GPP SAE effort. WiMAX Forum NWG Release 1 considers internetworking to 3GPP Release 7 through Interworking WLAN architecture[3] [TS 23.234 and TS 33.234] defined in 3GPP. WiMAX is attached to 3GPP as untrusted non-3GPP access. There are two access mechanisms described: direct access and WiMAX 3GPP IP access corresponding to scenario 2 and 3, respectively.

Reference framework is illustrated in Fig. 13.4. In order to have access to IP services, MS has to perform the initial network entry procedure via Wa interface to 3GPP AAA server. 3GPP AAA server creates the PMIP keys and distributes these keys to the HA and PMIP client residing in WiMAX network. WiMAX network is linked to WAG (WLAN Access Gateway) through Wn interface. The WAG acts as a

[3] The WLAN based on IEEE 802.11b/g/a and ETSI BRAN HiPERLAN2 are considered as untrusted non-3GPP access in 3GPP. 3GPP defines evolved Packet Data Gateway to introduce WLAN 3GPP IP Access to establish connectivity with external IP networks, such as 3G operator networks, corporate Intranets, or the Internet via the 3GPP system.

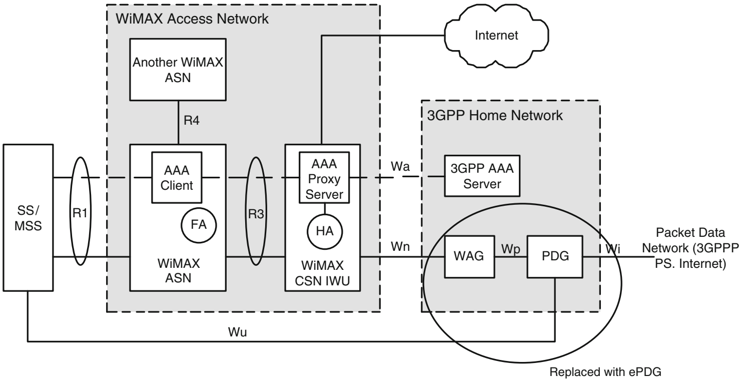


Fig. 13.4 WiMAX-3GGP interworking for nonroaming case

gateway to route the packets coming from WiMAX network. It enforces the routing of packets through the appropriate PDG. Wu is the tunnel between MS and PDG for user data packet transmission.

13.1.5 WiMAX Internetworking with DSL

Internetworking to DSL may be established at different stages of the end-to-end DSL network depicted in Fig. 13.5. DSL deployments are based on either PPP over Ethernet as link protocol between BRAS and TE (terminal equipment) or IP over Ethernet model. The former is selected for pure data transmission and the latter is selected for data, voice, and video services. PPP provides IP address acquisition and control of IP link. PPP packets need to be encapsulated into PPPoE frames on top of Ethernet for the point-to-point connectivity. IP over Ethernet on the other hand relies on DHCP for IP address configuration and identification of TE.

Fixed WiMAX based on IEEE 802.16-2004 can be integrated to DSL through V interface as seen in the figure. This way DSL reach can be extended over WiMAX with ETH-CS in BS. Mobile WiMAX introduces also A10 interface interworking in addition to V interface interworking. A10 interface interworking enables terminals beyond WiMAX access system to connect to DSL services. Interworking Unit (IWU) introduced for Mobile WiMAX integration as seen in Fig. 13.5 is a point to translate the WiMAX R5 interface to A10 interface of DSL network.

For A10 interface interworking, IWU is responsible to perform the following functions for respective convergence layers:

- PPPoETH with ETH-CS: (a) Receives PPP packet from the Layer 2 Tunneling Protocol Network Server (LNS) of DSL NSP with Layer 2 Tunneling Protocol (L2TP). (b) Removes L2TP header and adds correct PPPoE header. (c) Finds matched MAC address for TE and adds Ethernet frame header. (d) Finds matched

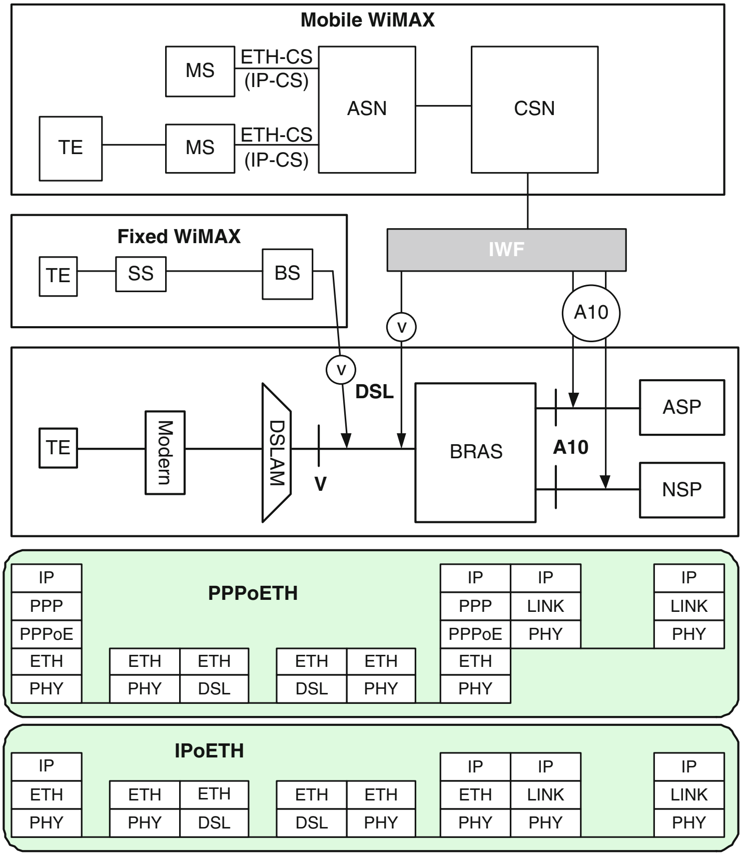


Fig. 13.5 DSL reference architecture: T interface is between terminal equipment and DSL modem in customer premises. V interface is Ethernet aggregation in the access network. A10 interface is between the access network and service providers. This interface connects to Application Service Provider via A10 ASP or Network Service Provider via A10 NSP

GRE key for TE. (e) Adds the GRE and IP header for packet. (f) Sends the IP packet to ASN-GW/CSN.

- IPoETH with ETH-CS: (a) Receives IP packet from the LNS of DSL NSP. (b) Finds matched MAC address for TE and adds Ethernet frame header. (c) Finds matched GRE key for TE. (d) Adds the GRE and IP header for packet. (e) Sends the IP packet to ASN-GW/CSN.
- IP-CS: (a) Receives IP packet from DSL NSP/ASP. (b) Relays the packet to ASN-GW/CSN.
- Proxy PPP with IP-CS: (a) Receives PPP over L2TP packet from the LNS of DSL NSP. (b) Removes L2TP and PPP header. (c) Finds matched GRE key for TE. (d) Adds the GRE and IP header for packet. (e) Sends the IP packet to ASN-GW/CSN.

For V interface, IWU is responsible for the following:

- PPPoETH with ETH-CS: (a) Receives Ethernet frame from the BRAS. (b) Finds matched GRE key for TE. (c) Adds the GRE and IP header for packet. (d) Relays the IP packet to ASN-GW/CSN.
- IPoETH with ETH-CS: (a) Receives Ethernet frame from the BRAS. (b) Finds matched GRE key for TE. (c) Adds the GRE and IP header for packet. (d) Relays the IP packet to ASN-GW/CSN.
- Proxy PPP with IP-CS: (a) Receives PPPoE frame from the BRAS. (b) Removes PPPoE and PPP header. (c) Finds matched GRE key for TE. (d) Adds the GRE and IP header for packet. (e) Sends the IP packet to ASN-GW/CSN.

13.1.6 GAN: Generic Access Network (formerly UMA)

Mobile operators target to seize the new opportunity arising due to abundance in dual-mode handsets[4] that support GSM (Global System for Mobile Communications) and Wi-Fi as seen in Fig. 13.6. The 3GPP GAN standards evolved from an UMA (Unlicensed Mobile Access) specification specify the handling of secure connectivity, registration, and transmission in order to let subscribers make calls inside the home over subscriber's unlicensed spectrum technologies (WLANs) to leverage VoIP over broadband service.

A GAN-enabled handset after detecting the unlicensed wireless network contacts GAN Controller (GANC) over the broadband IP access network to be authenticated and authorized to access 3GPP voice and data services via the unlicensed wireless network. The GAN framework introduces Up interface between MS and GAN controller in order to create a tunnel to MS to build security and control over unsecure wireless access. GANC has interfaces to MSC, SGSN via A and Gb interfaces, respectively, in addition to Wm interface to AAA.

Lately, also GAN network introduces Iu interfaces in order to support femtocells as seen in Fig. 13.7. The femtocell proposals considered three options: Iu-b over IP, SIP/IMS, Iu over IP to RAN Gateway. Traditionally femtocell proposals considered lu-b interface to RNC, which exists between 3G RNCs and macro 3G Node Bs. But due to the basic design of RNCs that are optimized to support low number of very high capacity macro base stations and proprietary Iu-b interfaces, this method is fell out of favor. SIP/IMS method proposes to provide services from SIP core network when a handset is connected to a femtocell. SIP network may be favored less than RAN Gateway since operator needs to replicate the services in macronetwork as well as in SIP core.

The RAN Gateway-based method proposes a Iu[5] over IP interface from femtocell to a RAN gateway (GANC), which implements lu-CS and lu-PS interfaces to MSC

[4] "In-Stat forecasts that consumers will use more than 66 million dual-mode handsets by 2009."

[5] Iu-h is being considered as new interface for LTE to connect home Node B (LTE femtocell) to LTE network.

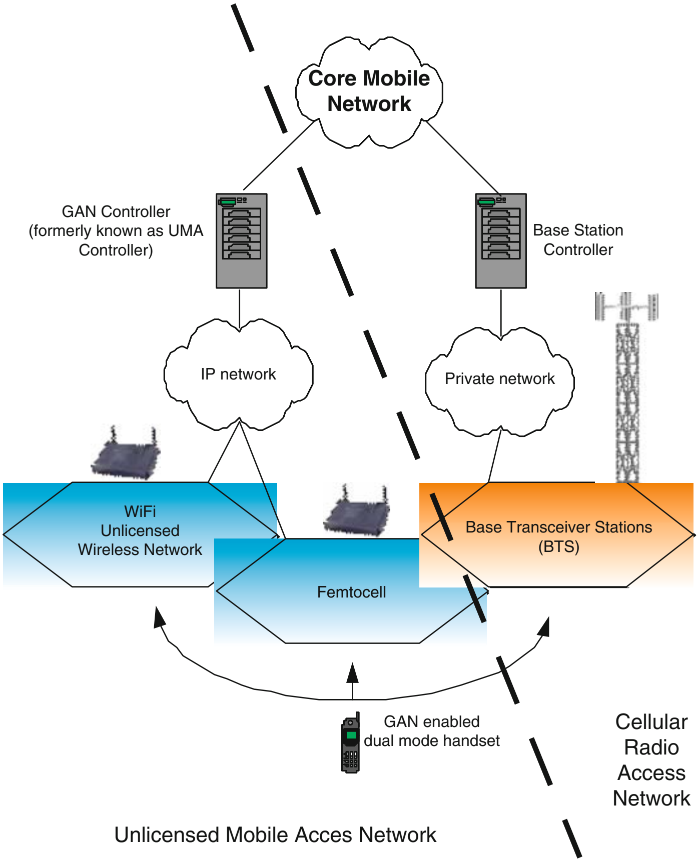

Core Mobile Network
GAN Controller (formerly known as UMA Controller)
Base Station Controller
IP network
Private network
WiFi Unlicensed Wireless Network
Femtocell
Base Transceiver Stations (BTS)
GAN enabled dual mode handset
Cellular Radio Access Network
Unlicensed Mobile Acces Network

Fig. 13.6 UMA

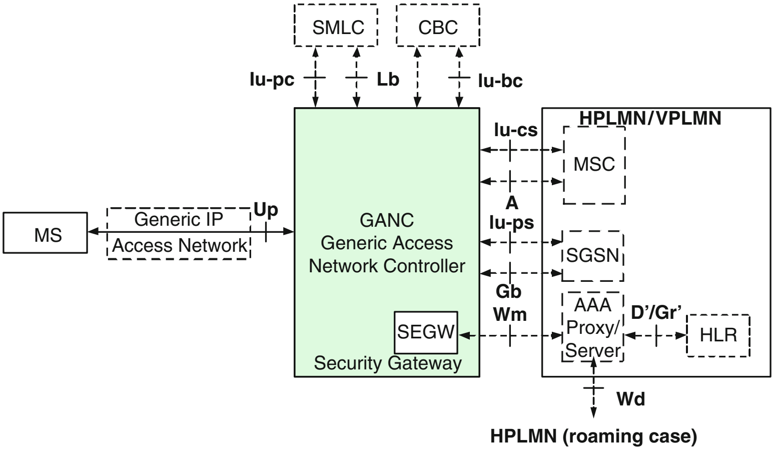


Fig. 13.7 UMA (source: TS 43318-800)

and SGSN. The RAN Gateway lets mobile operators use the standard interfaces and brings low initial cost of deployment with flat-IP architecture in which RNC functions are moved to femtocell for scaling issues. The GAN specification intends to extend the services toward FMC and provides migration path toward an all-IP converged infrastructure.

13.1.7 Seamlessness with IEEE 802.21

Multiradio devices will be abundant with advances in device and access technology. As we see, each access technology is designing its own way of integration to other access technologies. A standard way of integration that is agnostic to the underlying access technology is needed. IEEE 802.21 (media-independent handover services) framework aims to develop a common functionality for handover to bridge these access networks via a media independent handover in order to fulfill homogeneous or heterogeneous handovers. Homogeneous handover is a term defined for horizontal handovers within an access network and heterogenous handover is a term defined for vertical handovers between access networks for global mobility as seen in Fig. 13.8.

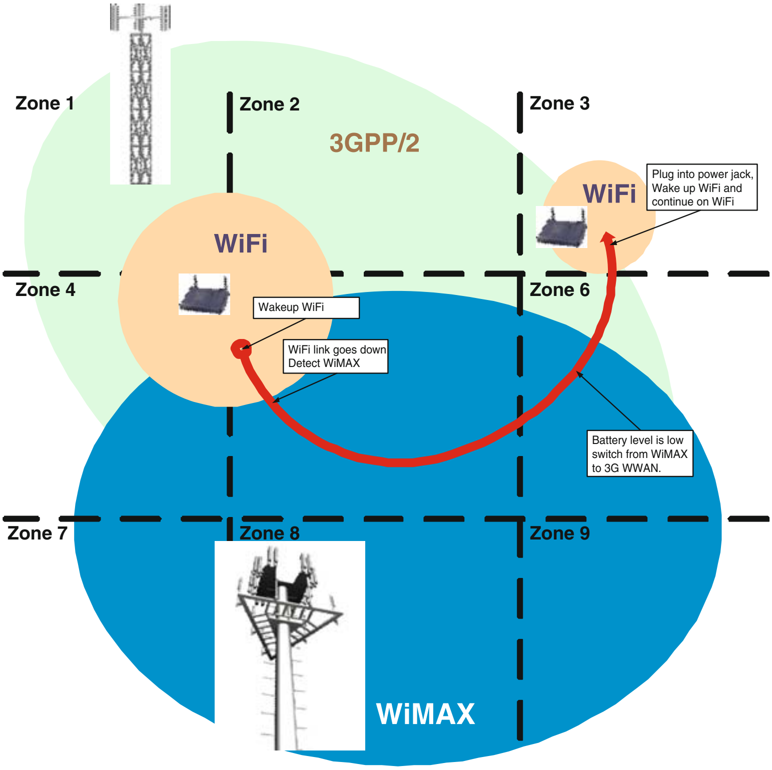

Zone 1
Zone 2
Zone 3
3GPP/2
WiFi
Plug into power jack, Wake up WiFi and continue on WiFi
WiFi
Zone 4
Zone 6
Wakeup WiFi
WiFi link goes down Detect WiMAX
Battery level is low switch from WiMAX to 3G WWAN.
Zone 7
Zone 8
Zone 9
WiMAX

Fig. 13.8 IEEE 802.21 Scenario

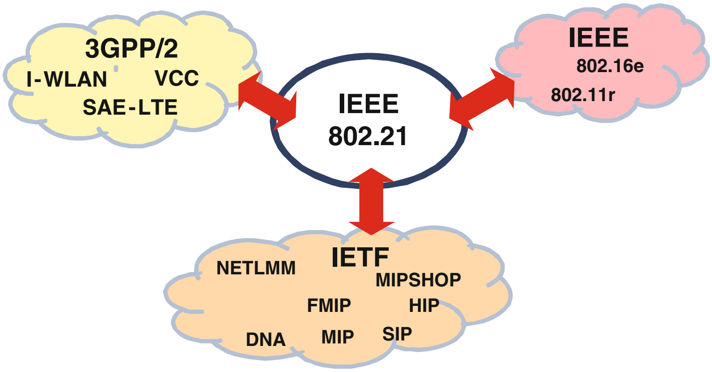


Fig. 13.9 IEEE 802.21 framework

IEEE 802.21 desired to bring a common handover solution for multiradio operation for better interworking of various technologies as seen in Fig. 13.9. Interworking includes handover within 802 services such as 802.3, 802.11, and 802.16, etc. It is also desired to extend 802.21 handover toward cellular or wired services.

The handover is defined as a three-step process: handover initiation, handover preparation, and handover execution. The scope of IEEE 802.21 is to cover handover initiation and handover preparation where functionalities include network discovery, selection, and handover negotiation in the former and layer 2 and 3 connectivity in the latter. On the other hand, remaining functionalities such as handover signaling, context transfer, and packet reception fall into handover execution.

IEEE 802.21 is being designed to develop smart L2 triggers, media-independent information service, and handover messages in order to provide optimum network selection, seamless roaming, and lower power operation for multiradio devices. L2 triggers consist of link up/down events, link parameters change/going down events, and network-initiated events for load balancing and operator preferences. IEEE 802.21 also proposes a media-independent information service that maintains a global network map that contains list of available networks (802.11/16/22, GSM, UMTS, LTE, etc.) with neighborhood information and available higher layer services such as ISP, MMS, etc. Handover decision is a co-operative feature with respect to triggers and neighbor information. Information element may contain the following in TLV or XML form:

- List of networks available
- Geo-location of Point of Attachment (PoA)
- Operator ID
- Roaming partners
- Cost indication for service/network usage
- Security
- Quality of Service
- PoA capabilities such as Emergency Services, IMS, etc.
- Vendor specific IEs

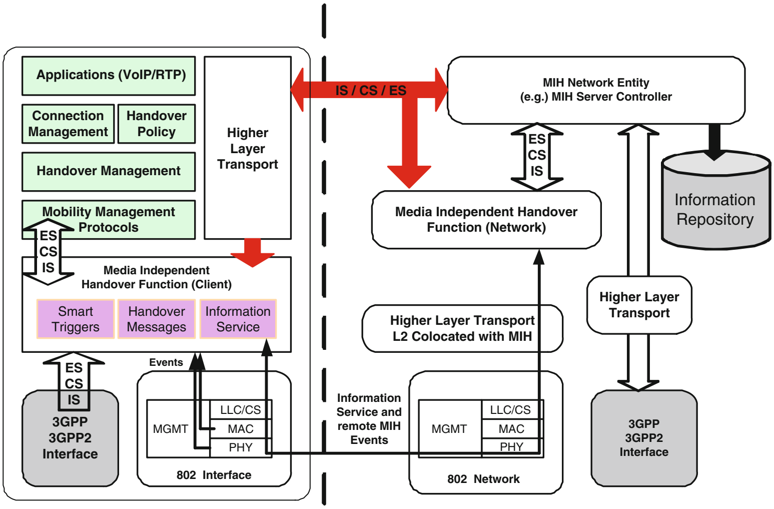


Fig. 13.10 IEEE 802.21 services: IEEE 802.21 has certain media-specific amendments to 802.11u (describes interwork with external networks), 802.16g, 3GPP-SAE, and IETF. Abbreviations in the figure are as follows: IS stands for Information Service, CS stands for Command Service, ES stands for Event Service, and LLC stands for Logical Link Controller

Figure 13.10 depicts the system architecture where MIH sublayer has been included above Convergence Sublayer in protocol stack. In WiMAX, MIH capability is conveyed with DCD message via MIH Capability Support Indication. In 802.11 it is conveyed with beacons and DHCP.

13.2 Service Convergence

Users want services that can be delivered at any time, over any device with any content and in one bill. They want services agnostic to network infrastructure, so that they can choose those services whenever and wherever they want. This is termed as "service convergence," which moves customers to a set of compelling services that are not tied to any one network or device. This is the driving force behind IP Multimedia Subsystem (IMS) framework and Policy and Charging Control (PCC) framework and Open Mobile Alliance.[6]

[6] "OMA is an industry forum formed in June 2002 in order to develop market driven, interoperable mobile service enablers. OMA targets to consolidate all specification activities in the service enabler space. OMA creates interoperable mobile data service enablers that work across devices, service providers, operators, networks, and geographies. Toward that end, OMA will develop test specifications, encourage third party tool development, and conduct test activities that allow vendors to test their implementations. Significant new work in OMA is leading to the development of mobile service enablers in areas such as Device Management, Push-to-talk Over Cellular, Mobile Broadcast, and more (source: `http://www.openmobilealliance.org/`).

13.2.1 One PCC

Convergence of next generation networks aims to deliver a unified end-user experience. To achieve this a central QoS control must be provisioned to ensure that the user experience provisioned is achieved over differing strands of communication. Unlike DiffServ and IntServ of traditional networks, QoS signaling for next generation networks requires an entity to perform application signaling, which do not necessarily travel on the same logical path as the actual data transfer itself. This entity is responsible to perform QoS authorization, QoS mapping, and provisioning the resultant QoS policy. The policy entity is also responsible to maintain end-to-end control where operation may span across multiple networks, carriers, and service providers. Convergence on the policy control entity's roles and functions varies depending on the standards but WiMAX is being designed to integrate into 3GPP and 3GPP2 PCC framework.[7] The PCC is originated in 3GPP and becoming a universal framework for dynamic QoS and charging management. The PCC framework introduce a means to enforce QoS and charging policies uniformly across all network elements for IMS-enriched IP-based 4G networks.

WiMAX interfaces to PCC of 3GPP/2 are illustrated in Fig. 13.11 with corresponding reference points. WiMAX defines IP-CAN bearer binding, which associates PCC rules to WiMAX service flows. It is an association with MS and an IP network and establishes right after IP address acquisition and terminates after IP address is released by MS. Before discussing the IP-CAN bearer binding, let us introduce the entities in the PCC framework:

- *PCRF:* Policy and Charging Rules Function encompasses policy control decision and flow-based charging control functionalities. It authorizes QoS rules and instructs data plane on how to proceed with the underlying data traffic. PCRF binds AF session and applicable PCC rules to an IP-CAN session.
- *PDF:* Policy Distribution Function hides the distributed nature of enforcement points from PCRF. It is connected to anchor SFA in the ASN via PCC-R3-P interface. The DIAMETER-based connection to PCRF is with Gx interface in 3GPP and Ty interface in 3GPP2. It acts as a distribution point to proxy Gx (Ty) messages to C-PCEF or to A-PCEF.
- *CDF:* Charging Distribution Function also hides the distributed nature of enforcement points from the OCS (Online Charging System) and supports SFA

[7] Also following is ongoing standardization on QoS control in next generation networks:

- ITU-T (International Telecommunication Union) is standardizing Next Generation Network under the umbrella of the Global Standards Initiative (NGN-GSI). ITU-T NGN Release 1 introduces Resource and Admission Control Functions (RACF).
- ETSI TISPAN (Telecoms and Internet converged Services and Protocols for Advanced Networks) introduces Resource and Admission Control Subsystems (RACS).
- PCMM (PacketCable MultiMedia) is a CableLabs-led initiative and introduces Policy Server (PS).
- MSF (Multiservice Switching Forum) introduces Bandwidth Manager (BM).

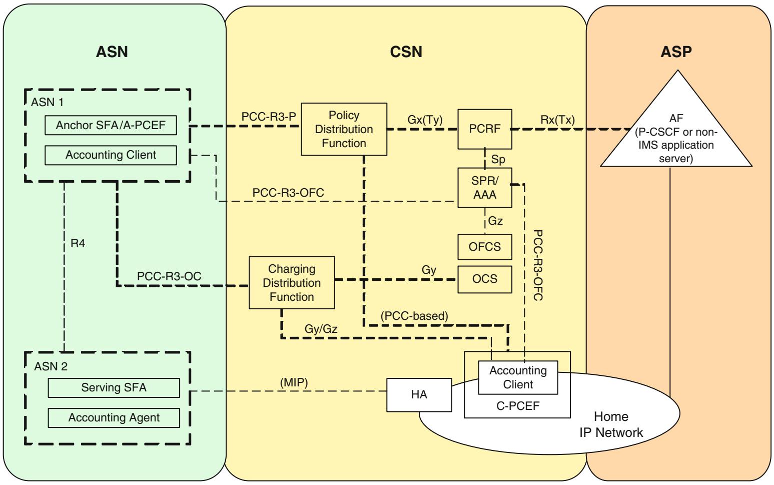


Fig. 13.11 PCC architecture

relocation. The CDF is connected to Accounting Client in the ASN via PCC-R3-OC interface and it is connected to OCS through Gy interface.

- *C-PCEF:* Policy and Charging Enforcement Point in the CSN is an optional enforcement point for the IP-level policies and/or charging. C-PCEF is on the data plane and communicates to PCC, CDF, and AAA for policies, online charging, and offline charging, respectively.
- *A-PCEF:* Policy and Charging Enforcement Point in the ASN is an enforcement point of PCC rules and/or charging in the ASN. Anchor SFA is called A-PCEF and it is responsible for relaying QoS-related policies to Serving SFA, relaying charging information to Accounting Client, and relaying PCC Service data flow template to Anchor DPF.
- *OCS/OFCS:* Online Charging System is responsible to provide the credit information to the PCEF via the *Gy* reference point (DIAMETER based) if online charging is applicable. Offline Charging System (OFCS) on the other hand uses Gz reference point for collecting offline charging record.
- *AF:* Application Function is an element in the service plane that requires dynamic policy and QoS control over the traffic plane behavior. The P-CSCF performs AF functionality in an IMS network.
- *AAA:* AAA proxies offline accounting messages from accounting client to the OFCS via Gz reference point (DIAMETER is desired).
- *SPR:* Subscriber Policy Register provides subscriber-specific data to the PCRF in order to assist in evaluating policy decisions.

Binding mechanism in PCC first performs session binding where PCRF associates the AF session and PCC rules to an IP-CAN with respect to the packet data network the user is accessing and the user-related information (IP address, identity, etc.). AF session and its service flows are described by a Service Data Flow ID

(SDFID) in WiMAX AF where a SDFID may be associated to one or more Packet Data Flow IDs (PDFIDs) in CSN. For instance, a SDFID for a video conference application may have two PDFIDs for voice and video flows. A PDFID is mapped to one or two SFID(s) (if traffic is bidirectional) in the A-PCEF of ASN and a SFID maps to IP-CAN bearer in the PCC. PCRF performs PCC rule authorization in order to associate a QoS class identifier to each IP-CAN bearer.

IP-CAN establishment starts after IP address acquisition in which A-PCEF sends a PCC rule request to PDF with QoS profile obtained from AAA. PDF communicates this to home PCRF, which checks SPR if subscriber's subscription-related information is missing (SPR also updates if there is any change.). Home PCRF makes the authorization and policy decision and sends it to the PDF over visited PCRF if roaming along with default charging mode. PDF acknowledges the IP-CAN session establishment to A-PCEF, which installs PCC rules and applies policy enforcement. A-PCEF relays the QoS policy and service flow level policies to serving SFA and SFM. The Serving SFA (Anchor DPF) enforces the IP policies, and SFM (BS) enforces service flow level QoS policies over the air. A-PCEF also initiates the charging by relaying the charging policy to collocated accounting client, which relays the charging policy to accounting agent for enforcements. A-PCEF notifies AAA when accounting starts and AAA notifies OFCS. IP-CAN session is not terminated when MS is in idle and sleep mode but when IP address is released. Termination can be initiated either by MS, BS, A-PCEF, or PCRF. Termination is notified to AF through PDF and PCRF and accounting stop is sent to OFCS through AAA. During A-PCEF mobility, new A-PCEF indicates the IP-CAN session modification to PDF only that notifies the old A-PCEF. As a result, new A-PCEF starts the accounting and old A-PCEF stops the accounting. Also, if SFA moves, A-PCEF notifies the PDF through IP-CAN session modification procedure.

13.2.2 One IMS

New communication culture based on new IP-based presence-enabled services is dictating anywhere, anytime, and on any device connectivity. Operators need to build and expand their services while taking advantage of the new communication culture as well as their expanded reach with the network convergence phenomena. In the previous section, we talk about common QoS control for users over differing standards. The same phenomena are required for common services since new services require interoperability between operators, networks, and devices.

IMS ensures interoperability of new services with operators, networks, and devices – which is built on ensuring support for global standards. IMS is a session control entity managing sessions that could come from different access networks. Interworking of these access networks to a common IMS with common PCC avails common applications with horizontal service offering. Traditionally, the network structure is very complex since each service is built independently from another and implementations of each layer must be built for every service. IMS on the other hand provides an end-to-end framework where any service can be built upon and horizontal layered architecture enables reuse of common functions and enablers by multiple applications.

IMS requires an IMS client to access IMS applications in the mobile subscriber. IMS client conforms all the MS procedures, IMS call control, and SIP extensions and applications. It needs to be compatible with OMA Device Management Protocol.

IMS session can be activated anytime after discovering the P-CSCF address(es). For instance in WiMAX, in CMIP scenario, CMIP client sends a DHCP Option with a SIP Server Option to retrieve P-CSCF address(es) or a list of FQDNs (fully qualified domain names) of P-CSCF(s). For non-CMIP scenarios, client may also utilize SIP Server option of DHCP Request message sent during IP address configuration. In roaming case, P-CSCF can be assigned by the home NSP or visited NSP via AAA signaling. When authentication request is forwarded to visited AAA from NAS. V-AAA forwards the access-request message with addresses of P-CSCFs in VCSN. H-AAA decides for the proper P-CSCF and appends the assigned P-CSCF address(es) or a list of FQDN in the access-accept message.

PCC correlation is depicted in Fig. 13.12 for WiMAX IMS architecture. After IP address acquisition, MS establishes IP-CAN session with PCC as described in the previous section. IP-CAN establishment is followed with P-CSCF discovery to enable IMS registration. If MS initiates IMS Session setup with IMS service, P-CSCF requests QoS authorization from PCRF for the bearer of the SIP session via Rx interface. PCRF initiates service flow establihsment/activation for the bearer of the IMS session.

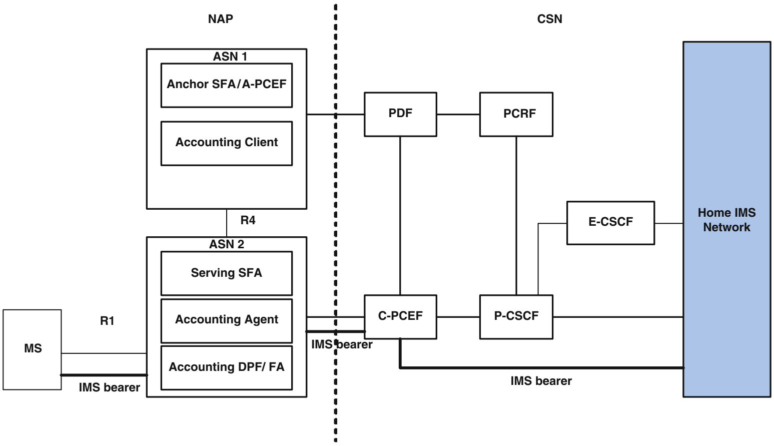


Fig. 13.12 Nonroaming reference model

13.3 Device Convergence

Finally, it is important to mention the trends in device convergence. Typically, a device is used for a single purpose and it has limitation to support other functions. As a result, when a new service is introduced it comes with a new device. What is needed is unifying devices that can access services in a similar and easy way. Device convergence in the handset is started with dual-mode devices in smart phones where multiple radio interfaces enables access over different networks. This is fortified with SIP that enables applications to traverse different IP networks. In this section, we discuss three trends: OHA, OBSAI, and ATCA.

OHA is formed by technology and mobile companies to deliver a richer, less expensive, and better mobile experience. OHA and Google have introduced Android mobile platform in 2008 at Mobile World Congress. The Android platform is a software stack for mobile devices, which has an operating system, middleware, and key applications. Other big vendors on mobile platform are Symbian, Nokia, Microsoft, and Apple. OBSAI is also an industry initiative, which is formed to create a complete set of open interface specifications related to the base station subsystem. The initiative tries to reduce the cost of the base station by utilizing the Commercial Off-the-Shelf (COTS) components and industry-led open specifications. This lets manufacturers utilize their resources more on value-added enhancements within the base station. Hence, not only operators benefit from greater innovation but also end-users.

ATCA is introduced in 2005 by PCI Industrial Computer Manufacturers Group (PICMG, www.picmg.org) as a follow-on to the widely used Compact PCI architecture of PICMG. ATCA specifications define an open and modular architecture for telecommunications equipment using COTS components. An equipment based on ATCA standards[8] is a plug-and-play architecture as seen in Fig. 13.13 where components include chassis, fabric switches, blades, shelves, high-availability operating software, and FCAPS[9] (Fault Configuration Accounting Performance and Security) solutions. As the largest specification effort in PICMG's history with more than 100 companies, ATCA has created a large ecosystem that independently delivers these components.

[8] "The ATCA specifications are defined as PICMG 3.X:

- *PICMG 3.0* defines the core specification including architecture mechanicals, power, system management, fabric connectors, and Base interface (10/100/1000 Base-T).
- *PICMG 3.1* defines the specification for Ethernet and Fibre Channel Fabric Interface.
- *PICMG 3.2* defines the InfiniBand Fabric interface.
- *PICMG 3.3* defines the StarFabric/Advanced Switching interface.
- *PICMG 3.4* defines the specification for PCI Express and Advanced Switching Fabric interface."

[9] "Similar to OAM&P (Operations, Administration, Maintenance, and Provisioning) is a term that is used to describe the collection of disciplines as well as software packages that is used to track these things."

Fig. 13.13 ATCA chassis and blades

13.4 More Coverage with IEEE 802.16j

Starting from this section we continue to introduce upcoming OFDMA-based technologies. The IEEE has started in 2006 to work on "Mobile Multihop Relay (MMR)" under 802.16j framework. IEEE 802.16j focuses enhancements to OFDMA physical layer and MAC layer to enable operation of a wireless relay station in order to enhance coverage, throughput, and system capacity with multihop relay capabilities and functionalities of interoperable relay stations.

One of the main operational expenditure drivers in mobile networks today is the requirement to connect each base station directly to the network. Typically, this is fulfilled by installing a cable or microwave connection at each base station. Currently, UMTS/HSPA base stations use E-1 or T-1 links, each capable of transmitting about 2Mbps. However, this backhaul link capacity will increase with OFDMA air link. Also consider the situations where the demand for capacity is ad hoc, then installing a base station would be wasteful.

The basic idea behind MMR is to allow WiMAX base stations that do not have a backhaul connection to communicate with base stations that do with some portion of the air link bandwidth. It is a simple and elegant way of extending network. This operation requires following modifications:

- A new frame structure to support in-band relaying
- Extension to security sublayer
- Enhanced MAC layer to handle bandwidth requests, handover, and packet delivery, etc.

The proposed modifications apply to new BS and RS (Relay Station) and work with existing 802.16e MS. IEEE 802.16j considers two types of relay modes as seen in Fig. 13.14: transparent relay and nontransparent relay.

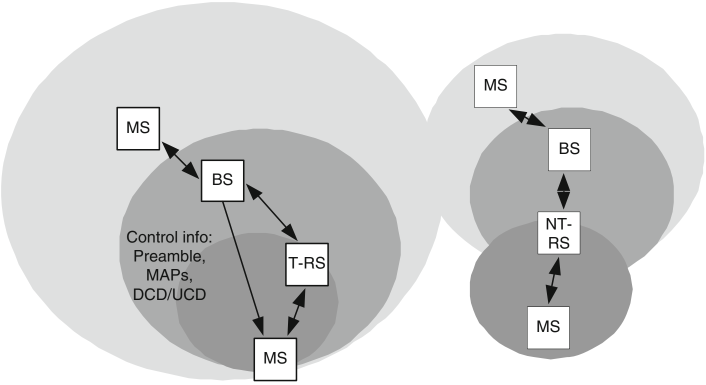


Fig. 13.14 Relay modes for IEEE 802.16j

13.4.1 Transparent Relay Mode

Transparent Relay Stations (T-RS) are typically considered for the situations where base stations' control information can reach the MS but data can be relayed through the transparent relay nodes. Control information such as preamble, FCH, MAP, and DCD/UCD messages are sent with the robust modulation and also can be fortified with multiple antenna systems. But, downlink coverage does not incur the same uplink coverage. As a result, MS can talk to BS through Transparent RS in the uplink.

OFDMA frame is divided into four zones as seen in Fig. 13.15: DL access zone, DL transparent zone, UL access zone, and UL relay zone. DL access zone hosts the downlink control information as well as downlink bursts. DL transparent zone is dedicated for RS to communicate to another RS or MS. UL access zone has a resource allocation for MS to communicate to RS, and UL relay zone is for RS to relay the uplink bursts of MSs to the BS. Both uplink zones have a ranging subchannel.

Transparent relay mode has centralized scheduling in the BS. BS generates the MAPs and RSs have to conform those MAPs when generating their MAPs.

13.4.2 Nontransparent Relay Mode

Nontransparent Relay Stations (NT-RS) work like a base station and they are allowed to transmit downlink control information. This type of operation is suitable when BS cannot reach the MS. Figure 13.16 illustrates the frame structures for nontransparent relay mode. From the figure we can see that the BS addresses MSs and RSs, respectively, in access or relay zones of downlink and uplink portion of the frame. In DL relay zone, BS transmits the MAP, which is specific to RSs, and

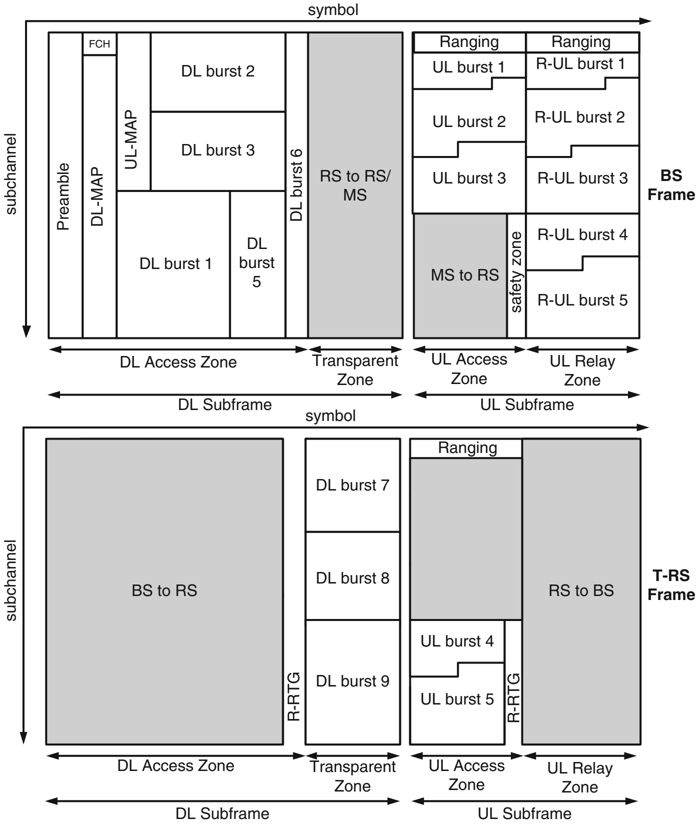


Fig. 13.15 Transparent relay mode frame structure

in the UL relay zone, RSs relay the uplink bursts to the BS according to the MAP. Note that MAPs in the DL access zone of the NT-RS frame are created by the RS and independent from the BS.

13.5 More Capacity with IEEE 802.16m

The 802.16m (WiMAX-m) amendment is to provide an advanced air interface for operation in licensed bands in order to meet the cellular layer requirements of ITU-R and IMT-advanced next generation mobile networks. It also provides support for legacy WirelessMAN-OFDMA (WiMAX-e) equipment. The backward compatibility requirements are as follows:

- IEEE 802.16m MS shall operate with legacy BS, at a level of performance equivalent to that of a legacy MS.

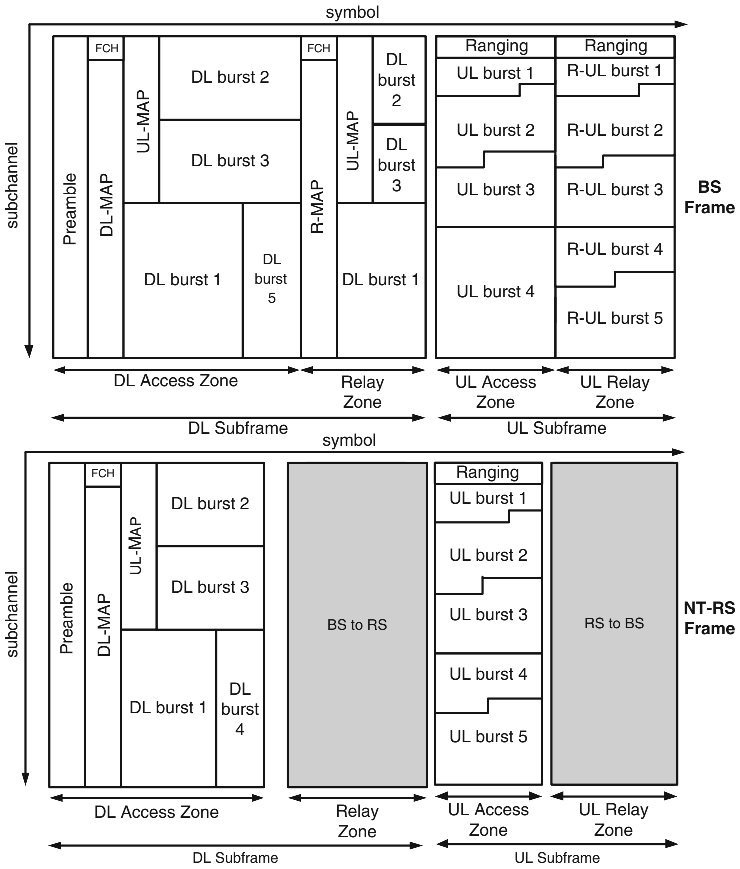


Fig. 13.16 Nontransparent relay mode frame structure

- IEEE 802.16m and the WirelessMAN-OFDMA shall operate on the same RF carrier, with the same channel bandwidth, and should be able to operate on the same RF carrier with different channel bandwidths.
- IEEE 802.16m should support a mix of IEEE 802.16m and legacy MSs on the same frequency. System performance improves with the fraction of IEEE 802.16m MSs.
- IEEE 802.16m shall support handover of a legacy MS from/to a legacy BS.

The IEEE 802.16m system should be able to use spectrum flexibly to provide TDD and H/FDD duplexing modes and should be capable of coexisting with other IMT-advanced and IMT2000 technologies. Under ideal conditions peak data rates and average throughput of data only system are specified in Tables 13.1 and 13.2.

The data latency envisioned is 10 ms and idle state to active state transition is required to be 100 ms at maximum. Handover interruption time, which is the time MS does not get packet from BS, is 30 and 100 ms at maximum for intrafrequency

Table 13.1 Normalized peak data rate

Requirement type	MIMO configuration	Normalized rate (bps/Hz)
Baseline DL	2 × 2	8.0
Baseline UL	1 × 2	2.8
Target DL	4 × 4	15.0
Target UL	2 × 4	5.6

Table 13.2 Absolute throughput of data only system

Metric	DL data (bits/s/Hz)	UL data (bits/s/Hz)
Average throughput	0.26	0.13
Cell edge throughput	0.09	0.05

and interfrequency. It shall support at least 30 active VoIP users per MHz and per sector with 12.2 Kbps codec. Geo-location determination latency should be less than 30 s and the accuracy should be around 50–150 and 100–300 m for mobile-initiated location determination and network-initiated location determination, respectively. In addition to the required support for multihop relay and media-independent handover, IEEE 802.16m shall support codeployment and coexistence with other access technologies in the same band or adjacent band. Also, self-configuration and self-optimization support are required in order to enable plug and play installation of network nodes together with autonomous optimization.

13.5.1 Uplink

SC-FDMA has been proposed for uplink of IEEE 802.16m. Since legacy support is mandatory, there are basically TDM or FDM methods to multiplex single carrier users with 16/16e users as seen in Fig. 13.17.

The TDM (Time Division Multiplexing) separates the conventional OFDMA users and SC-FDMA users in time with a new SC-FDMA zone. The FDM (Frequency Division Multiplexing) separates those in frequency, which maps them into different subcarriers.

SC-FDMA has been proposed because of its lower PAPR or cubic metric (CM), which will be translated to improvements in power-amplifier efficiency and coverage area, and insensitive feature to the frequency offset.

On the other hand, OFDMA proposal states that PAPR might not be a major issue since with proper power control and scheduling PAPR can be addressed and only a small percentage of users at the cell edge need to use maximum power. Also, it is

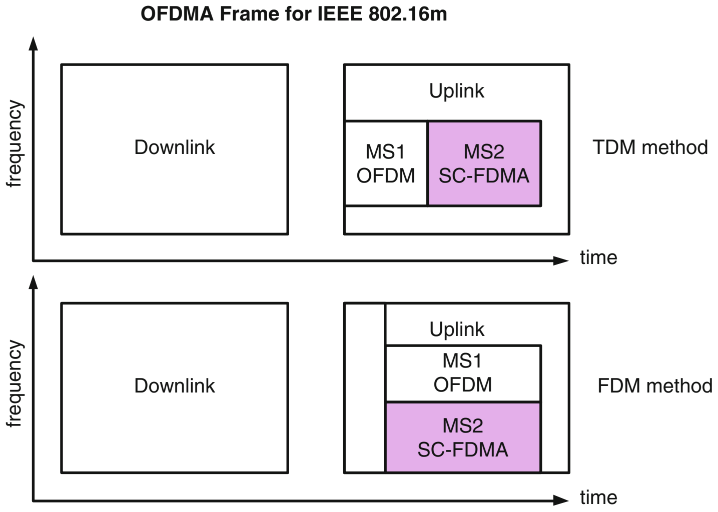


Fig. 13.17 SC-FDMA proposal

stated that SC-FDMA is disadvantageous since it requires higher receiver complexity, additional DFT processing in the mobile stations, and low flexibility in multiplexing uplink control and data channels.

13.5.2 Low-Latency Frame

IEEE 802.16m frame structure is required to be backward compatible with legacy system and need to be evolving from supporting legacy terminals to one which will be supporting an increasing number of 16m terminals. The coexistence of both legacy and 16m terminals will have no impact on the performance of legacy MSs.

The legacy support enforces certain requirements of legacy frame on 16m frame so that legacy MSs can successfully proceed to initialization, access, connected, and idle states. The following features must be present within the 16m frame structure:

- 16e Preamble every 5 ms
- FCH and DL-MAP after 16e preamble
- DCD/UCD transmitted every i^{th} frame
- Uplink ranging region every j^{th} frame
- Optional uplink ACK and CQICH regions

A superframe concept is introduced as in Fig. 13.18. Superframe is composed of 5-ms legacy frames in which with division 16m frame can be introduced with desired latency. This reduced latency will facilitate faster CQI feedback by MS to improve efficiency of high mobile users.

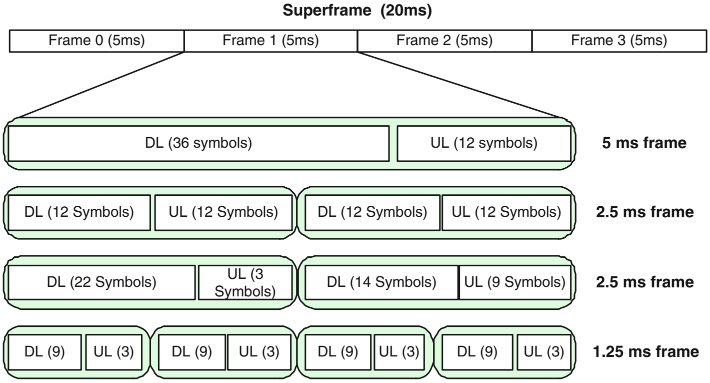


Fig. 13.18 Proposed IEEE 802.16m superframe structure

Figure 13.19 shows a possible evolution frame structure in which transition from legacy stations to 16m is depicted. In the beginning it is expected that legacy terminals will predominate. As a result, 16m zone enlarges dynamically as there are more 16m terminals.

13.6 More Access with IEEE 802.20

IEEE 802.20 or Mobile Broadband Wireless Access (MBWA) Working Group was established in December 2002 targeting low-cost, always-on, and truly mobile broadband wireless networks. The goals of IEEE 802.20 and IEEE 802.16e-2005, Mobile WiMAX, are the same but WiMAX has initially considered mobility in the 2–6 GHz band and 802.20 aims for operation below 3.5 GHz. A draft was approved in January 2006, and IEEE 802.20 ratification is expected in first half of 2008.

IEEE 802.20 introduces FDD and TDD modes of operation, and the Ultra Mobile Broadband (UMB) FDD mode described in the previous chapter is proposed to fulfill the FDD mode since UMB FDD and current MWBA FDD mode share common ancestry but UMB FDD has evolved faster with new changes and enhancements. Some enhancements include support for OFDMA control channel, optional CDMA Data Channel in the RL, optional rotational OFDM (DRCH mode only), and additional HARQ interlaces, etc.

MBWA TDD mode, on the other hand, has two associated variables in order to determine the time partitioning between forward and reverse links as seen in Fig. 13.20. The figure depicts 1:1 and 2:1 partitioning and HARQ structure for these is shown in Fig. 13.21.

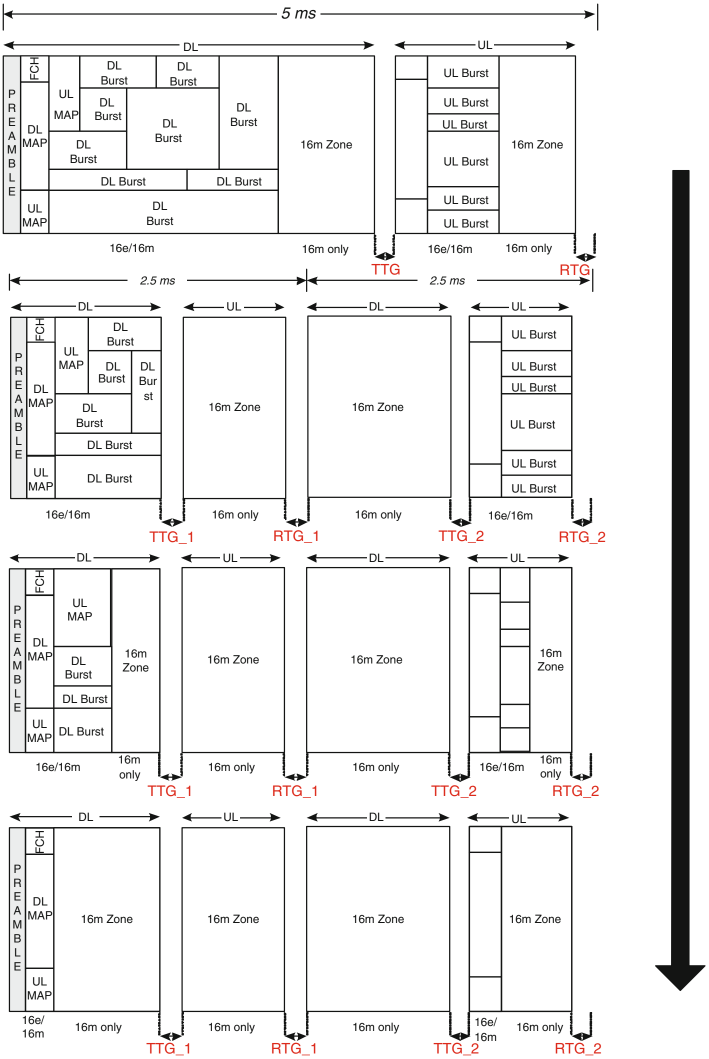
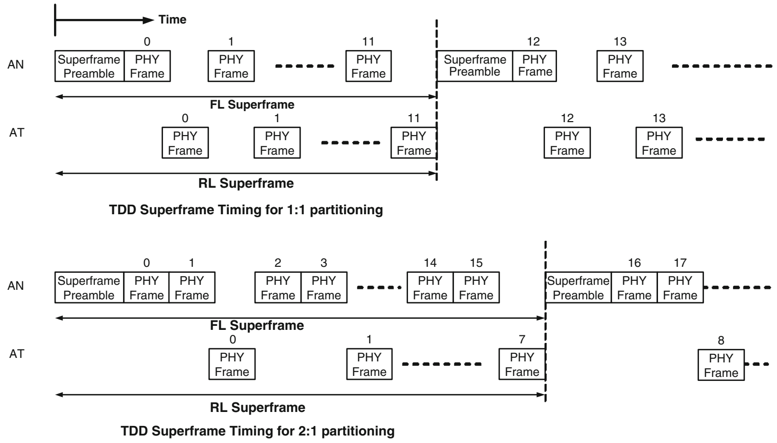


Fig. 13.19 Proposed IEEE 802.16m frame structure (© IEEE)

13.7 More Free Spectrum with IEEE 802.22

The IEEE 802.22[10] (cognitive radio) standard for Wireless Regional Area Networks (WRANs) (see Fig. 13.22 for comparison) is another technology that considers to adopt OFDMA technology in its air interface. The IEEE 802.22 working group is chartered with the deployment of the first air interface standard (MAC and PHY)

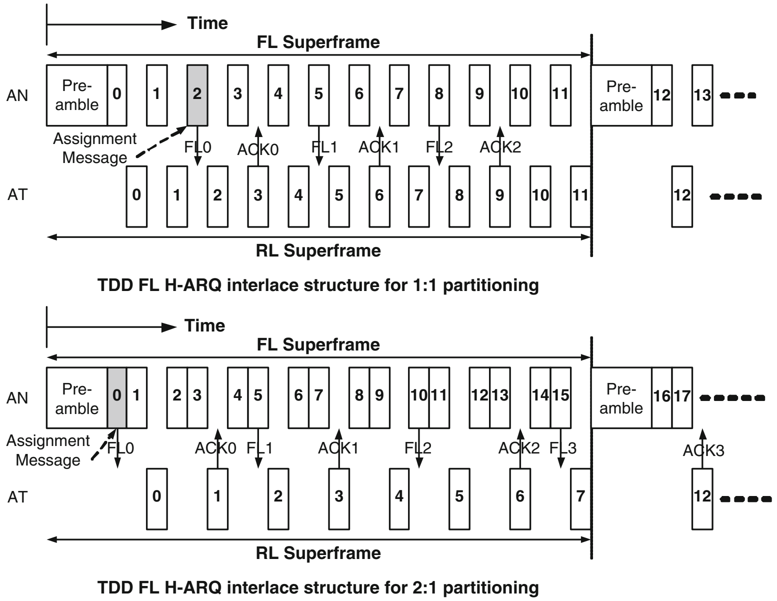
[10] http://www.ieee802.org/22/.



TDD Superframe Timing for 1:1 partitioning

TDD Superframe Timing for 2:1 partitioning

Fig. 13.20 TDD superframe



TDD FL H-ARQ interlace structure for 1:1 partitioning

TDD FL H-ARQ interlace structure for 2:1 partitioning

Fig. 13.21 TDD FL HARQ structure

based on cognitive radios to aim license-exempt operation in the fallow TV spectrum with no interference to TV receivers. The applications may include fixed broadband access technologies, content delivery to homes/offices with QoS support, etc.

Cognitive radio (CR) techniques have potential to offer unlicensed operation in licensed bands with intelligent spectrum sensing, allocation, and acquisition solutions in order not to interfere with other *incumbent* devices.

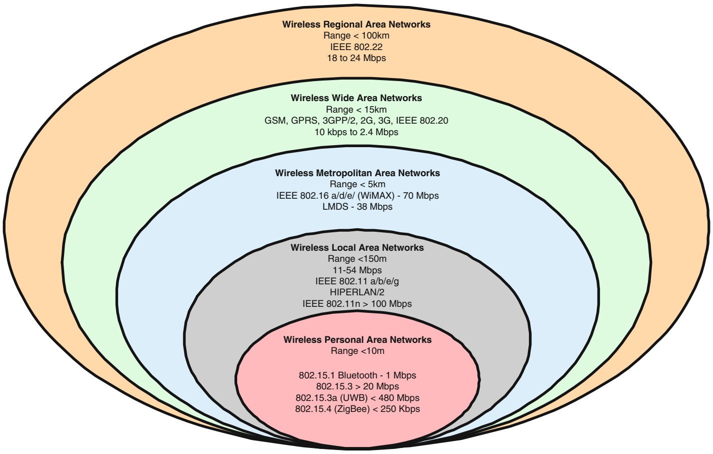

Wireless Regional Area Networks
Range < 100km
IEEE 802.22
18 to 24 Mbps
Wireless Wide Area Networks
Range < 15km
GSM, GPRS, 3GPP/2, 2G, 3G, IEEE 802.20
10 kbps to 2.4 Mbps
Wireless Metropolitan Area Networks
Range < 5km
IEEE 802.16 a/d/e/ (WiMAX) - 70 Mbps
LMDS - 38 Mbps
Wireless Local Area Networks
Range <150m
11-54 Mbps
IEEE 802.11 a/b/e/g
HIPERLAN/2
IEEE 802.11n > 100 Mbps
Wireless Personal Area Networks
Range <10m
802.15.1 Bluetooth - 1 Mbps
802.15.3 > 20 Mbps
802.15.3a (UWB) < 480 Mbps
802.15.4 (ZigBee) < 250 Kbps

Fig. 13.22 Comparison of standards

The FCC has introduced Notice of Proposed Rule Making (NPRM) for TV band[11] in May 2004 to propose to allow unlicensed radios in TV bands with no harmful interference to incumbents through CR techniques. The TV bands are of particular interest since they are of low frequency, and so can provide a wide coverage with lots of bandwidth available.

In the USA, the White Space Coalition is formed by big players such as Google, Microsoft, Hewlett Packard, and Dell in order to utilize “White Space” between analogue television transmissions – space becoming more available with the conversion to digital to provide high-speed wireless broadband. In November 2008, FCC officials approved a plan for white space wireless broadband. Previously, IEEE has formed the 802.22 group for Wireless Regional Area Networks (WRANs) in November 2004. The proposed IEEE 802.22 system considers fixed access as in fixed WiMAX but in wider area. Also, IEEE 802.22 base station requires distributed sensing capacity and self-coexistence feature to protect the incumbent receivers as well as other 802.22 operations.

13.7.1 IEEE 802.22 Air Interface

The proposed operation detects the vacancies in time and frequency as in Fig. 13.23. As a result, two-dimensional agility offered by OFDMA is required to blank out the

[11] “TV stations in USA operate in 54–72, 76–88, 174–216, and 470–806 MHz with 6-MHz wide spectrum. Also wireless microphones, and Private Land and Commercial Mobile Radio Services (PLMRS/CMRS) including Public Safety are also allowed. There is also debate to extend the operation to 41–910 MHz to meet the international regulatory requirements with 6, 7, and 8-MHz bandwidths. ”

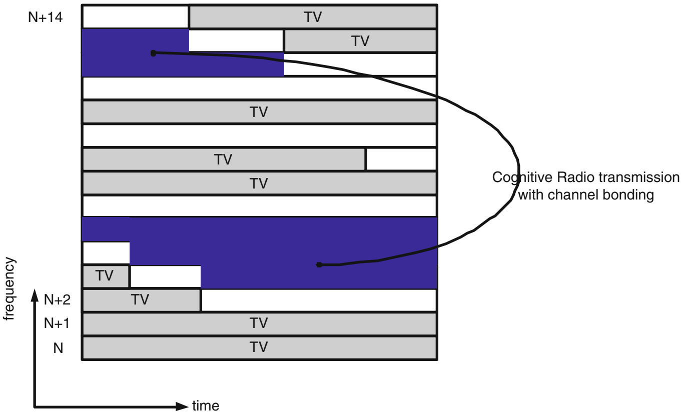


Fig. 13.23 Operation of IEEE 802.22: The WRAN device operation is restricted to have at least three contiguous vacant TV channels (6K FFT for 18 MHz) since US grade-A TV allocation requires at least two empty channels between TV channels

certain time and frequencies. Also, adaptive modulation and coding is required to accommodate subscribers spread over wide range to provide either more bandwidth to subscribers close to base station or more robust transmission for those away from the base station.

The proposed OFDMA PHY considers 2,000 carriers in one TV channel of 6-MHz wide with long cylic prefix, which is around 40 μs due to long delay spread in large terrain. Modulation schemes supported are proposed to be QPSK, 16QAM, and 64QAM with 1/2, 3/4, and 2/3 coding rates in order to scale speed up to 19 Mbps per TV channel through 48 subchannels. The 802.22 also requires channel bonding scheme in which contiguous and noncontiguous vacant channels can be used to increase the bandwidth with fixed subcarrier spacing.

The MAC operation defines two frames: superframe and frame. The MAC superframe and frame in the current 802.22 draft are illustrated in Fig. 13.24. A superframe contains a preamble and superframe control header sent in each TV channel up to three if detected vacant. This header gives the necessary information such as bonding, etc. The superframe may contain multiple MAC frames, which may span multiple channels. As can be seen in Fig. 13.24 when incumbent operation starts at channel $n+2$, CR-based radio stops transmitting in TV channel $n+1$, thus bonding is allowed in two channels only. The Channel Detection Time, which is the time during which an incumbent withstands the interference, is specified by the Dynamic Frequency Selection (DFS) model, ordered by the FCC for 5-GHz band.

The MAC frame operates in TDD mode and has adjustable downlink and uplink subframe ratio. The operation is similar to TDD mode of IEEE 802.16e but there could be possible coexistence intervals. Also in uplink, there is an *urgent coexistence situation* region in addition to ranging, bandwidth request, and data regions.

The spectrum management requires in-band and out-of-band measurements. BS coordinates and consolidates the readings sent by the subscriber devices. There

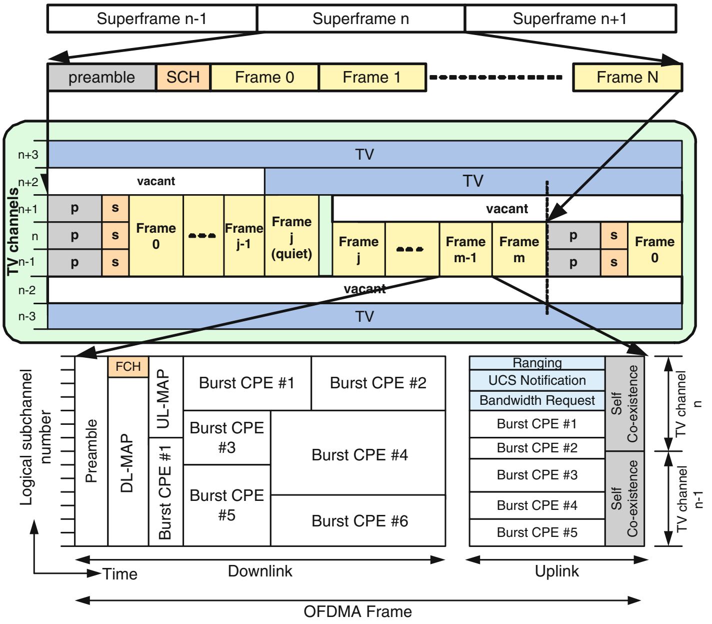


Fig. 13.24 IEEE 802.22 MAC superframe and frame structures

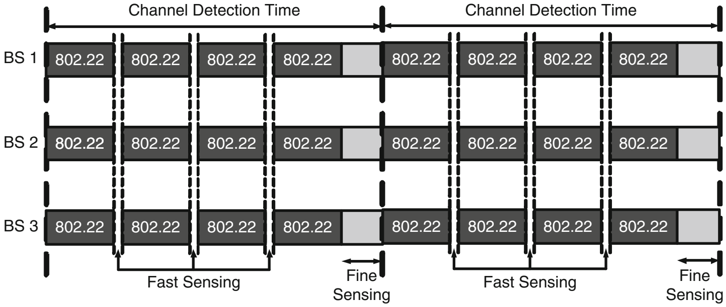


Fig. 13.25 Sensing in IEEE 802.22

are two antennas in a subscriber: one is directional for data transmission and the other is omnidirectional for sensing and measurements. For in-band measurement, subscribers and the BS silence the transmission for less than 1ms to detect energy periodically in order to perform fast sensing as seen in Fig. 13.25. If energy is detected, then fine sensing is performed in the order of milliseconds (∼25 ms). These sensing eras are synchronized among BS to address coexistence of multiple 802.22 BSs. Typically, −116, −94, and −107 dBm thresholds are considered, respectively, for Digital TV, Analog TV, and wireless microphones. The MAC maintains a spectrum usage table, which is updated by the system provider.

13.8 Summary

The migration to converged networks is real and accelerating. The convergence not only supports anytime anywhere mobility, but also it has as its heart OFDMA technology. In this chapter, we introduced the drivers for convergence and covered the upcoming OFDMA technologies. Key highlights of this chapter are as follows:

- Fixed-mobile convergence targets network convergence, device convergence, and service convergence for unified communication at any time anywhere with any device.
- Interworking of access technologies is the current driving trend to enable network convergence. Interworking among 3GPP/2, WiMAX, DSL, and WiFi is being designed to support interworking solutions ranging from loosely coupled to integrated. Media-independent handover is considered in IEEE 802.21 framework to enable technology agnostic handover.
- Common Policy and Charging Control and IP Multimedia Subsystem provide converged services. Device convergence on the other hand is happening in handsets with open standards and in network equipment with standardized solutions such as ATCA.
- New amendments to existing technologies and future standards are considering OFDMA as the underlying air interface. Within WiMAX context, mobile multihop relay structure is being designed in IEEE 802.16j workgroup to extend coverage of a base station with wireless mobile/fixed relay stations. Also, IEEE 802.16m amendment to existing 802.16 standard considers an OFDMA-based air interface to meet the IMT-advanced criteria for next generation networks.
- IEEE 802.20 is another OFDMA-based standard for mobile broadband wireless access. IEEE 802.20 inherits similar OFDMA functionalities as Flash-OFDM and UMB.
- IEEE 802.22 is a cognitive radio-based access technology to utilize vacant TV channels in order to provide wireless regional access. IEEE 802.22 considers OFDMA air interface together with a cognitive-based spectrum sensing.

In brief, there are number of reasons why there is a need for converged networks – the vast majority of which point to efficiency, scalability, and connectivity. In today's industry with finite resources in terms of R&D engineers and funding, witnessing a parade of merging standards is required rather than "duelling"[12] ones.

References

1. Jette, A., "UMBFDD Draft Technology Overview," IEEE 802.20 Working Group document, March 2007. `http://grouper.ieee.org/groups/802/20`.
2. Rappaport, T.S., *Wireless Communications*, Prentice Hall, 1996.

[12] Belongs to Arun Sarin, CEO of Vodafone, told during his keynote speech in Mobile World Congress, February 2008.

3. Cordeiro, C., Challapali, K., Birru, D., Shankar NS., "IEEE 802.22: An Introduction to the First Wireless Standard based on Cognitive Radios," *Journal of Communications*, vol.1, no.1, April 2006.
4. Federal Communications Commision (FCC), *Spectrum Policy Task Force*, ET Docket no. 02-135, November 15, 2002.
5. Mitola, J., et al., "Cognitive Radios: Making Software Radios more Personal," *IEEE Personal Communications*, vol.6, no.4, August 1999.
6. Haykin, S., "Cognitive Radio: Brain-Empowered Wireless Communications," *IEEE JSAC*, vol.23, no.2, February 2005.
7. Federal Communications Commission (FCC), "Notice of Proposed Rule Making," ET Docket no. 04-113, May 25, 2004.
8. DARPA ATO, Next Generation (XG) Program, http://www.darpa.mil/ato/programs/xg/.
9. Federal Communications Commission (FCC), "Revision of Parts 2 and 15 of the Commissions Rules to Permit Unlicensed National Information Infrastructure (U-NII) Devices in the 5GHz Band," ET Docket No. 03-122, November 18, 2003.
10. Chouinard, G., "WRAN Reference Model," IEEE 802.22, May 2005. http://www.ieee802.org/22/.
11. Recommendation ITU-R M.1645, "Framework and overall objectives of the future development of IMT 2000 and systems beyond IMT 2000".
12. Gupta, V., Williams, M. G., Jonhston, D.J., Barber, P., Ohba, Y., "IEEE 802.21 Overview of Standard for Media Independent Handover Services," http://www.802.org/21.
13. "UMA Technology Overview," http://www.umatechnology.org/overview/index.htm.
14. Cudak, M., "IEEE 802.16m System Requirements," http://wirelessman.org/tgm.
15. Kang, J., Boariu, A., Li, S., "Proposal for Incorporating Single-Carrier FDMA into 802.16m," IEEE C802.16m-08/100, January 2008. http://wirelessman.org/tgm.
16. Tee, A., Ma, J., Fong, M. H., Jia, M., Sivanesan, K., Yu, D., Kim, S-Y., "IEEE 802.16m Multiple Access Techniques," IEEE S602.16m-08/112r1, January 2008. http://wirelessman.org/tgm.
17. Mc Daid, C., "Overview and Comparison of QoS Control in Next Generation Networks," Palo Wireless. http://www.palowireless.com/3g/qos.asp.
18. 3GPP Standard TS 23.203 v7.1.0, "Policy and Charging Control Architecture," December 2006. http://www.3gpp.org.
19. 3GPP2 Draft X.S0013-012-0 v0.21.0, "Service Based Bearer Control," April 2006. http://www.3gpp2.org.
20. WiMAX Forum Release 1.5 Draft, "Policy and Charging Control," February 2008. http://www.wimaxforum.org.
21. In-Stat, "Wireless IP Phones Drive Future VoIP Markets," August 2005.
22. FCC Cognitive Radio Technologies Proceedings, http://www.fcc.gov/oet/cognitiveradio.

List of Figures

List of Tables

Glossary

3GPP	Third Generation Partnership Project
3GPP2	Third Generation Partnership Project 2
A-MIMO	Adaptive multiple input multiple output
AA	Anchor authenticator
AAA	Authentication, authorization, and accounting
AAS	Adaptive antenna system (also advanced antenna system)
AASN	Anchor ASN
AC	Admission control
ACK	Acknowledge
ADPF	anchor data path function
AES	Advanced encryption standard
AF	Application function
AG	Absolute grant
aGW	E-UTRAN access gateway
AK	Authorization key
AK SN	Derivation from PMK and PMK2 SN
AKA	Authentiction and key agreement
AM	Authorization module
AMC	Adaptive modulation and coding
AMS	Adaptive MIMO switching
APC	Anchor paging controller
APCF	Anchor paging controller function
API	Application program interface
AR	Access router
ARIB	Association of radio industries and businesses
ARQ	Automatic repeat request
AS	Authentication server or access stratum
ASN	Access service network
ASP	Application service provider
BCE	Binding cache entry
BE	Best effort

BGCF	Breakout gateway control function
BRAS	Broadband remote access server
BS	Base station
BSID	Base station identifier
BU	Binding update
CAC	Connection admission control
CC	Chase combining (also convolutional code)
CCI	Co-channel interference
CCM	Counter with cipher-block chaining message authentication code
CCoA	Collocated care of address
CDF	Charging distribution function
CDF	Cumulative distribution function
CDM	Code division multiplex
CDMA	Code division multiple access
CDMA2000	Third generation code division multiple access radio technology
CID	Connection identifier
CINR	Carrier to interference + noise ratio
CMAC	Cipher-based message authentication code
CMIP	Client mobile IP
CoA	Care of address
COA	Change of authority
COS	Class of service
CP	Control plane or cyclic prefix
CQI	Channel quality indicator
CS	Convergence sublayer
CSN	Connectivity service network
CSTD	Cyclic shift transmit diversity
CTC	Convolutional turbo code
CUI	Chargeable user identity
CWTS	China wireless telecommunication standard group
DAD	Duplicate address detection
DECT	European digital enhanced cordless telecommunications
DHCP	Dynamic host configuration protocol
diffserv	Differentiated services
DL	Down link
DNS	Domain name service
DOCSIS	Data over cable service interface specification
DoS	Denial of service
DP	Decision point or data path
DPCCH	Downlink physical control channel
DRC	Data rate control
DSC	Data source control
DSL	Digital subscriber line
DSLAM	Digital subscriber link access multiplexer
DVB	Digital video broadcast

E-AGCH	E-DCH absolute grant channel
E-DCH	Enhanced data channel
E-DPDCH	E-DCH dedicated physical data control channel
E-HICH	E DCH HARQ acknowledgement indicator channel
E-RGCH	E-DCH relative grant channel
E-UTRA	Evolved universal terrestrial radio access
E-UTRAN	Evolved universal terrestrial radio access network
E2E	End-to-end
E911	US emergency services
EAP	Extensible authentication protocol
EAP-AKA	EAP authentication and key agreement
EAP-MD5	EAP – message digest 5
EAP-PSK	EAP – pre-shared key
EAP-SIM	EAP subscriber identity module
EAP-TLS	EAP with TLS
EESM	Exponential effective SIR mapping
EIRP	Effective isotropic radiated power
EMSK	Extended master session key
eNB	E-UTRAN nodeB
ERT-VR	Extended real-time variable rate
ertPS	Extended real-time polling service
ETSI	European Telecommunications Standards Institute
EUI-64	Extended unique identifier (64-bit)
EVDO	Evolution data optimized or evolution data only
EVDV	Evolution data-voice
FA	Foreign agent
FBSS	Fast base station switching
FCAPS	Fault configuration accounting performance and security
FCH	Frame control header
FDD	Frequency division duplex
FFT	Fast Fourier transform
FQDN	Fully qualified domain name
FRD	Fast router discovery
FTP	File transfer protocol
FUSC	Fully used subcarrier
FWA	Fixed wireless access
GF	Galois field
GPRS	General packet radio services
GRE	Generic routing encapsulation
GSA	Group Security Association
GSM	Global system for mobile communication
GW	Gateway
HA	Home agent
HARQ	Hybrid automatic repeat request
HHO	Hard hand-off

HLA	Hot-line application
HLD	Hot-line device
HMAC	Keyed-hashing for message authentication code
HO	Hand-off or hand over
HO ID	Hand-off identifier
HoA	MS home address
Hotspot	Public location where WLAN services have been deployed
HRPD	High-rate packet data
HS-DPCCH	High-speed dedicated physical control channel
HS-DSCH	High-speed downlink shared channel
HS-SCCH	High-speed shared control channel
HSDPA	High-speed downlink data packet access
HSPA	High-speed packet access
HSUPA	High-speed uplink data packet access
HTTP	Hyper-text transfer protocol
I-WLAN	Interworking with wireless LANs
IANA	Internet assigned numbers authority
IBS	Integrated base stations
ICMPv6	Internet control message protocol for IPv6
IE	Information elements
IEEE	Institute of Electrical and Electronics Engineers
IEEE 802.3	IEEE standard specification for Ethernet
IEFT	Internet Engineering Task Force
IFFT	Inverse fast Fourier transform
IID	Interface identifier
IK	Integrity key
IKEv2	Internet key exchange protocol version 2
IMS	IP multimedia subsystem
IMSI	International mobile subscriber identity
IP	Internet protocol
IPsec	IP security
IPv4	Internet protocol version 4
IPv6	Internet protocol version 6
IR	Incremental redundancy
ISF	Initial service flow
ISI	Intersymbol interference
ISM	Industrial, scientific, and medical bands
IWF	Internetworking function
IWG	Interworking gateway
IWU	Internetworking Unit
L1	Layer 1 (physical layer)
L2	Layer 2 (data link layer)
L3	Layer 3 (network layer)
LBS	Location-based services
LDPC	Low-density parity check

LE	License-exempt deployments
LMDS	Local multipoint distribution system
LOS	Line of sight
LPF	Local policy function
LR	Location register MSID, BSID
LSB	Least-significant byte
LTE	Long-term evolution
MAC	Medium access control
MAI	Multiple access interference
MAN	Metropolitan area network
MAP	Media access protocol
MBMS	Multimedia broadcast/multicast service
MBS	Multicast and broadcast service
MCC	Mobile country code
MDHO	Macrodiversity hand over
MIMO	Multiple input multiple output
MIP	Mobile IP
MIP6	Mobile IP version 6
MLD	Maximum likelihood symbol detection
MM	Mobility management
MMS	Multimedia message service
MMSE	Minimum mean-squared error
MNC	Mobile network operator code
MN_HOA	Allow-MN-HA assignment
MPLS	Multi protocol label switching
MS	Mobile station
MSID	Mobile station identifier
MSK	Master session key
MSO	Multiservices operator
NA	Neighbor advertisements
NACK	Not acknowledge
NAI	Network access identifier
NAP	Network access provider
NAPT	Network address port translation
NAS	Network access server or Nonaccess stratum
NAT	Network address translation
NLOS	Non–line-of-sight
NMS	Network management system
NRM	Network reference model
NRT-VR	Non–real-time variable rate
nrtPS	Non–real-time polling service
NS	Neighbor solicitation
NSP	Network service provider
NUD	Neighbor unreachability detection
OAM	Operations and maintenance

OFDM	Orthogonal frequency division multiplex
OFDMA	Orthogonal frequency division multiple access
OTA	Over-the-air
OUI	Organization unique identifier
P-CSCF	Proxy-call session control function
PA	Paging agent
PBX	Private branch exchange
PC	Paging controller
PDCP	Packet data convergence protocol
PDFID	Packet data flow ID
PDG	Packet data gateway
PDU	Packet data unit
PEAP	Protected EAP
PER	Packet error rate
PF	Policy function
PF	Proportional fair (scheduler)
PG	Paging group
PG ID	Paging group identifier
PHS	Packet header suppression (PHS)
PKM	Public key management
PMIP	Proxy-mobile IP
PMK	Pairwise master key
PMK2	Pairwise master key
PMN	Proxy mobile node
PoA	Point of attachment
PPAC	Prepaid accounting capability
PPC	Prepaid client
PPS	Prepaid server
Proxy-ARP	Proxy address resolution protocol
PS	Physical slot
PSK	Preshared key or phase shift keying
PSTN	Public switched telephone network
PtP	Peer to peer
PUSC	Partially used subcarrier
QAM	Quadrature amplitude modulation
QoS	Quality of service
QPSK	Quadrature phase shift keying
RA	Router advertisement or reverse activity
RAB	Reverse-link activity bit
RADIUS	Remote access dial in user service
RG	Relative grant
RLC	Radio link control
RNC	Radio network controller
RO	Route optimization
RP	Reference point

RPC	Reverse power control
RR	Resource-reservation or round Robin
RRA	Radio resource agent
RRC	Radio resource controller
RRI	Reverse rate indicator
RRM	Radio resource management
RS	Router solicitation
RS	Reed-Solomon coding
RSVP	Resource reservation protocol
RT-VR	Real-time variable rate
RTG	Receive/transmit transition gap
rtPS	Real-time polling service
RUIM	Removable user identity module
S-CSCF	Serving-call session control function
S-OFDMA	Scalable orthogonal frequency division multiple access
SA	Security association
SAE	System architecture evolution
SCI	Spare capacity indicator
SDFID	Service data flow ID
SDMA	Space (or spatial) division (or diversity) multiple access
SDU	Service data unit
SF	Spreading factor
SFA	Service flow authorization
SFID	Service flow ID
SFM	Service flow management
SFN	Single frequency network
SGSN	Serving GPRS support node
SHO	Soft hand-off
SI	Subscriber identity
SII	System information identity or service identity information
SIM	Subscriber identity module
SIMO	Single input multiple output (antenna)
SINR	Signal to interference + noise ratio
SISO	Single input single output (antenna)
SLA	Service-level agreement
SM	Spatial multiplexing
SMS	Short message service
SMTP	Simple mail transport protocol
SNIR	Signal to noise + interference ratio
SNMP	Simple network management protocol
SNR	Signal to noise ratio
SS	Subscriber station
SS7	Signaling system 7
SSL	Secure sockets layer
STBC	Space–time block code

STC	Space–time coding
SUBC	Subscriber credentials
T1	Committee T1
TBS	Target BS
TCH	Traffic channel
TCP	Transmission control protocol
TD-CDMA	Time division code division multiple access
TD-SCDMA	Time division synchronous code division multiple access
TDD	Time division duplex
TDM	Time division multiplex
TE	Terminal equipment
TEK	Traffic encryption key
TFRI	Transport format-related information
TIA	Telecommunications Industry Association
TLS	Transport layer security
TLV	Type length value
TTA	Telecommunications Technology Association
TTC	Telecommunication Technology Committee
TTG	Transmit/receive transition gap
TTI	Transmission time interval
TTLS	Tunneled TLS
TU	Typical urban (as in channel model)
U-NII	Unlicensed national information infrastructure
UDP	User datagram protocol
UDR	Usage data record
UE	User equipment
UGS	Unsolicited grant service
UICC	Universal integrated circuit card
UID	User identity
UL	Uplink
UMTS	Universal mobile telecommunications system
UP	User plane
USIM	Universal subscriber identity module
UTRAN LTE	UMTS terrestrial radio access network long-term evolution
V-AAA	Visited AAA proxy
V-MIMO	Virtual multiple input multiple output (antenna)
VLAN	Virtual LAN
VoIP	Voice over Internet protocol
VPN	Virtual private network
VSA	Vendor-specific attributes
VSF	Variable spreading factor
VSM	Vertical spatial multiplexing
WAG	WLAN access gateway
WAP	Wireless application protocol
WATSP	WiMAX ASN transport signaling protocols

WCDMA	Wideband code division multiple access
WEP	Wired equivalent privacy
Wi-Fi	Wireless fidelity
WiBro	Wireless broadband (service)
WiMAX	Worldwide interoperability for microwave access
WLAN	Wireless local area network
WPA	Wi-Fi protected access
WWAN	Wireless wide area network
X.509	ITU standard for digital public-key certificate issued by a CA

Index